

Il malinteso come struttura cognitiva

Una teoria formale dell'incomprensione produttiva

Proemio: Il malinteso come fatto primo

La filosofia del linguaggio e la teoria della comunicazione condividono un'assunzione che raramente viene messa in discussione esplicitamente: che il malinteso sia un fenomeno derivato, una deviazione dal caso normale della comprensione riuscita, un guasto nel meccanismo la cui funzione propria è la trasmissione fedele di significato. Su questa assunzione si costruiscono le teorie della comunicazione da Shannon e Weaver in poi: il rumore è il nemico, la ridondanza è lo strumento per combatterlo, e il successo comunicativo è il caso limite verso cui il sistema tende quando il rumore viene azzerato.

La tesi di questa monografia è che questa assunzione sia invertita. Il malinteso non è il caso degenerare della comunicazione — è il caso normale, quello strutturalmente inevitabile ogni volta che due sistemi cognitivi con apparati di rappresentazione non identici si scambiano segnali. La comprensione piena e simmetrica — due agenti che estraggono esattamente lo stesso contenuto da uno scambio comunicativo — è il caso limite irraggiungibile, non il caso di default. E la produttività della comunicazione — il fatto che lo scambio di segnali tra agenti diversi generi nuova conoscenza, nuovi concetti, nuove strutture — dipende precisamente dal fatto che il malinteso non è un guasto ma un generatore: è la fonte da cui emergono significati che nessuno dei due agenti possedeva prima dello scambio.

Questa tesi richiede una precisazione immediata. Non si sostiene che tutti i malintesi siano produttivi, né che la comprensione sia impossibile. Si sostiene che il malinteso ha una struttura formale propria — non è rumore uniforme ma distorsione sistematica nella direzione del sistema di rappresentazione del ricevente — e che questa struttura è la condizione di possibilità della comunicazione tra menti diverse, non il suo ostacolo. Un mondo di menti identiche non ha bisogno di comunicare; un mondo di menti radicalmente diverse non può comunicare. La comunicazione produttiva abita la zona intermedia — e quella zona è strutturata dal malinteso.

Il percorso si articola in cinque movimenti. Il primo costruisce il modello formale del malinteso come trasduzione tra sistemi di rappresentazione — non errore di trasmissione ma trasformazione sistematica del segnale attraverso il filtro del sistema di rappresentazione del ricevente. Il secondo analizza le massime conversazionali di Grice come protocollo di coordinazione che riduce il malinteso senza eliminarlo, e le conseguenze del fallimento di ciascuna massima. Il terzo tratta il malinteso creativo —

la produzione di significati nuovi attraverso la trasformazione del segnale originale — come fonte primaria dell'evoluzione concettuale. Il quarto affronta il malinteso istituzionalizzato: le situazioni in cui le strutture sociali cristallizzano e perpetuano malintesi sistematici come meccanismi di potere. Il quinto chiude con il confronto tra malinteso umano e malinteso artificiale — il tipo specifico di distorsione che i modelli di linguaggio introducono nel processo comunicativo.

Capitolo I — Il modello formale: la comunicazione come trasduzione tra sistemi di rappresentazione

I.1 Il canale di Shannon e il suo limite

Il modello di Shannon della comunicazione descrive un canale che trasmette simboli da una sorgente a un destinatario, soggetto a rumore. La codifica del mittente e la decodifica del destinatario sono progettate per essere inverse l'una dell'altra: $E \circ D = \text{id}$ al limite del canale ideale. Il rumore è la deviazione dalla condizione ideale $E \circ D \neq \text{id}$, e il teorema di Shannon quantifica il massimo tasso di trasmissione fedele compatibile con un canale di capacità fissata.

Questo modello presuppone che mittente e destinatario condividano lo stesso codice — lo stesso sistema di rappresentazione — e che il problema comunicativo sia esclusivamente il problema del rumore fisico. Questa è la limitazione fondamentale per l'applicazione alla comunicazione cognitiva: due agenti cognitivi reali non condividono lo stesso sistema di rappresentazione. Il codice del mittente M_α e il codice del destinatario M_β sono sistemi con strutture diverse, e il processo di decodifica non è l'inverso della codifica ma una *trasduzione* — una trasformazione che mappa il segnale dallo spazio del mittente nello spazio del destinatario attraverso un processo che non è, in generale, invertibile.

I.2 La trasduzione come modello della comprensione

Si definisce formalmente la trasduzione comunicativa. Sia s il segnale trasmesso — una stringa su un alfabeto condiviso (il linguaggio naturale, o qualsiasi altro medium di comunicazione con una sintassi pubblica condivisa). Il mittente α produce s a partire dall'intenzione comunicativa i_α — il contenuto che intende trasmettere — attraverso la funzione di codifica:

$$s = \text{Enc}_\alpha(i_\alpha)$$

Il destinatario β riceve s e produce un'interpretazione i_β attraverso la funzione di decodifica:

$$i_\beta = \text{Dec}_\beta(s)$$

La comprensione perfetta richiederebbe $i_\beta = i_\alpha$. Il malinteso è $i_\beta \neq i_\alpha$. La struttura del malinteso è determinata dalla composizione $\text{Dec}_\beta \circ \text{Enc}_\alpha$ — la trasduzione complessiva che mappa l'intenzione del mittente nell'interpretazione del destinatario.

La trasduzione $\text{Dec}_\beta \circ \text{Enc}_\alpha$ non è in generale l'identità, per due ragioni distinte che meritano di essere tenute separate. La prima è la *lossy compression*: Enc_α comprime i_α nel segnale s , perdendo informazione che M_α considera ridondante o implicita. La seconda è la *distorsione di decodifica*: Dec_β non è l'inversa di Enc_α ma la migliore decodifica possibile di s nel sistema di rappresentazione M_β , che non coincide con M_α . Anche in assenza di rumore fisico sul canale, la trasduzione introduce distorsione sistematica.

I.3 La misura del malinteso: complessità condizionale e distorsione semantica

Si definisce la *distanza di malinteso* tra l'intenzione i_α del mittente e l'interpretazione i_β del destinatario come la complessità condizionale:

$$\text{MU}(\alpha \rightarrow \beta; i_\alpha) = \text{KC}(i_\alpha \mid i_\beta)$$

la lunghezza della descrizione più concisa di i_α dato i_β — quanto informazione aggiuntiva è necessaria per ricostruire l'intenzione originale a partire dall'interpretazione prodotta. $\text{MU} = 0$ indica comprensione perfetta: i_β è sufficiente a ricostruire i_α completamente. MU elevata indica malinteso profondo: l'interpretazione è così distante dall'intenzione che ricostruire quest'ultima richiede molta informazione aggiuntiva.

Si definisce separatamente la *distorsione semantica*:

$$\text{DS}(\alpha \rightarrow \beta; i_\alpha) = \text{KC}(i_\beta \mid i_\alpha)$$

la lunghezza della descrizione di i_β dato i_α — quanto l'interpretazione contiene di informativo che l'intenzione non aveva. $\text{DS} > 0$ indica che l'interpretazione del destinatario include struttura non presente nell'intenzione originale: il destinatario ha *aggiunto* significato al messaggio, non soltanto mal-ricevuto quello esistente.

Le due misure non sono in generale uguali. MU misura ciò che il mittente ha perso nella trasduzione; DS misura ciò che il destinatario ha aggiunto. Un malinteso con MU alta e DS bassa è un malinteso *impoveritivo*: il destinatario ha perso informazione senza aggiungerne. Un malinteso con MU bassa e DS alta è un malinteso *arricchitivo*: il destinatario ha aggiunto struttura all'intenzione originale. Un malinteso con entrambe alte è un malinteso *trasformativo*: il segnale ha subito una trasformazione radicale che ha perso la struttura originale e prodotto una struttura nuova.

I.4 La decomposizione del malinteso: struttura sistematica e rumore

Il malinteso non è uniforme — non è rumore casuale sulla trasduzione. Ha una struttura sistematica determinata dalla relazione tra i sistemi di rappresentazione M_α e M_β . Si decompone in due componenti.

La componente *sistematica* del malinteso è la distorsione introdotta dalla trasduzione stessa — la differenza tra i_α e $\text{Dec}_\beta(\text{Enc}_\alpha(i_\alpha))$ in assenza di rumore. Questa componente è deterministica data la coppia (M_α, M_β) e il segnale s : è lo stesso per ogni ricevente con sistema di rappresentazione M_β che riceva il segnale s . È prevedibile, analizzabile e — in linea di principio — compensabile se la struttura dei due sistemi di rappresentazione è nota.

La componente *stocastica* del malinteso è la varianza residua del processo di decodifica — le fluttuazioni che dipendono dallo stato contestuale del destinatario al momento della ricezione: il suo umore, la sua attenzione, la salienza delle sue aspettative in quel momento. Questa componente non è sistematica ma è comunque strutturata: ha la forma della varianza di generazione della memoria del Capitolo I della monografia precedente, modulata dalla cue rappresentata dal segnale s .

La distanza di malinteso totale si decompone come:

$$\text{MU}(\alpha \rightarrow \beta; i_\alpha) = \text{MU}_{\text{sys}}(\alpha, \beta) + \text{MU}_{\text{stoch}}(s, \theta_\beta)$$

dove MU_{sys} è la componente sistematica — funzione solo dei sistemi di rappresentazione — e MU_{stoch} è la componente stocastica — funzione del segnale specifico e dello stato del destinatario.

I.5 Il malinteso sistematico come proprietà della coppia di agenti

La componente sistematica $\text{MU}_{\text{sys}}(\alpha, \beta)$ è una proprietà della coppia di sistemi di rappresentazione (M_α, M_β) che non dipende dai contenuti specifici comunicati. Misura la distorsione strutturale intrinseca alla comunicazione tra i due agenti — il malinteso minimo inevitabile indipendentemente dal segnale.

MU_{sys} è asimmetrica: $\text{MU}_{\text{sys}}(\alpha, \beta) \neq \text{MU}_{\text{sys}}(\beta, \alpha)$ in generale. La comunicazione da α a β può produrre un malinteso sistematico diverso dalla comunicazione in direzione inversa, per le stesse ragioni asimmetriche che rendevano il costo di traduzione $\text{CT}(\alpha \rightarrow \beta) \neq \text{CT}(\beta \rightarrow \alpha)$ nella monografia precedente. Un esperto che comunica con un novizio non malintende il novizio quanto il novizio malintende l'esperto: la codifica dell'esperto usa primitive assenti nel sistema di rappresentazione del novizio, mentre la codifica del novizio usa primitive disponibili anche nell'esperto.

MU_{sys} ha la struttura di una semi-distanza sullo spazio dei sistemi di rappresentazione: soddisfa $\text{MU}_{\text{sys}}(\alpha, \alpha) = 0$ (un agente non si malintende con se stesso) e la disuguaglianza triangolare $\text{MU}_{\text{sys}}(\alpha, \gamma) \leq \text{MU}_{\text{sys}}(\alpha, \beta) + \text{MU}_{\text{sys}}(\beta, \gamma)$ (il

malinteso tra α e γ è al più la somma dei malintesi attraverso un intermediario β). Non è simmetrica, quindi non è una distanza nel senso metrico.

I.6 Le massime di Grice come protocollo di riduzione del malinteso

Paul Grice ha proposto che la comunicazione sia governata da un *principio di cooperazione* e da un insieme di massime che i partecipanti seguono — o fingono di seguire — per produrre comunicazione efficace. Le massime sono: quantità (dare la giusta quantità di informazione), qualità (dire solo ciò che si crede vero e per cui si ha evidenza), relazione (essere pertinenti), e modo (essere chiari, brevi, ordinati).

Nel quadro formale della trasduzione, le massime di Grice possono essere reinterpretate come vincoli sul processo di codifica Enc_α progettati per ridurre $MU_{sys}(\alpha, \beta)$. La massima di quantità vincola la lunghezza di $Enc_\alpha(i_\alpha)$: né troppo lunga (ridondanza che aumenta il rumore) né troppo breve (perdita di informazione necessaria alla decodifica). La massima di qualità vincola l'accuratezza di Enc_α : il segnale deve essere un'istanza dell'intenzione, non una distorsione di essa. La massima di relazione vincola la pertinenza di Enc_α : il segnale deve attivare nel destinatario le primitive di M_β più vicine all'intenzione, anziché primitive distanti. La massima di modo vincola la struttura di Enc_α : la forma del segnale deve minimizzare l'ambiguità nella decodifica Dec_β .

La formalizzazione precisa di ciascuna massima produce una funzione di costo che il codificatore ottimale minimizza. La massima di quantità produce il vincolo:

$$|Enc_\alpha(i_\alpha)| \approx R(D_{target}) / \log|\Sigma|$$

la lunghezza del segnale deve essere approssimativamente uguale alla lunghezza minima necessaria per garantire distorsione D_{target} sulla trasduzione — la lunghezza data dalla funzione di tasso-distorsione applicata alla comunicazione. Codifiche più brevi producono distorsione superiore a D_{target} (violazione per difetto); codifiche più lunghe producono ridondanza che può essere sfruttata dal rumore (violazione per eccesso).

I.7 Le implicature come malintesi costruttivi: il caso della comunicazione indiretta

Grice ha osservato che i partecipanti alla conversazione seguono le massime non solo letteralmente ma anche nella loro violazione apparente: quando un enunciato viola apparentemente una massima, il destinatario non interpreta la violazione come guasto comunicativo ma come segnale che il mittente intende comunicare qualcosa di diverso dal contenuto letterale. Questo è il meccanismo delle *implicature conversazionali*.

Nel quadro formale, un'implicatura è un caso di malinteso costruttivo — una deviazione intenzionale dal valore letterale $Enc_\alpha(i_\alpha)$ che il mittente usa per

trasmettere un'intenzione $i_{\alpha'}$ diversa dall'intenzione letterale i_{α} , sfruttando la struttura della trasduzione Dec_{β} . Il mittente produce un segnale $s = \text{Enc}_{\alpha}(i_{\alpha})$ — codifica dell'intenzione letterale — sapendo che $\text{Dec}_{\beta}(s)$ produrrà non i_{α} ma $i_{\alpha'}$ — l'interpretazione che il sistema di rappresentazione di β genera dalla violazione della massima.

Formalmente, un'implicatura riuscita richiede:

$$\text{Dec}_{\beta}(\text{Enc}_{\alpha}(i_{\alpha})) = i_{\alpha'} \quad \text{dove } i_{\alpha'} \neq i_{\alpha}$$

il mittente sfrutta la struttura della trasduzione per trasmettere $i_{\alpha'}$ usando il segnale di i_{α} . La condizione di riuscita è che il mittente abbia un modello sufficientemente accurato di Dec_{β} — che sappia quale interpretazione il segnale produrrà nel destinatario — e che quell'interpretazione sia quella che il mittente intende trasmettere.

Questo mostra che la comunicazione sofisticata non è solo utilizzo del codice condiviso ma *modellazione della trasduzione del destinatario* — utilizzo del malinteso sistematico come strumento. I mittenti esperti — i poeti, i retori, i diplomatici, i comici — sono esperti nella modellazione di Dec_{β} : sanno quale interpretazione il loro segnale produrrà in un destinatario con sistema di rappresentazione M_{β} , e costruiscono i segnali per produrre l'interpretazione desiderata attraverso la trasduzione, non nonostante di essa.

Capitolo II — Il malinteso creativo: come la distorsione genera significato nuovo

II.1 Il paradosso della traduzione

Ogni traduzione è un malinteso istituzionalizzato. Un traduttore che trasporta un testo da una lingua all'altra produce inevitabilmente $DS > 0$ — aggiunge struttura che il testo originale non possedeva — e $MU > 0$ — perde struttura che il testo originale possedeva. Non esiste traduzione con $MU = DS = 0$: la condizione equivarrebbe a un'identità tra i due sistemi linguistici, il che annullerebbe la necessità della traduzione stessa.

Il paradosso è che le traduzioni più apprezzate — quelle che hanno avuto il maggiore impatto sulla cultura ricevente — sono spesso le meno fedeli nel senso di bassa MU , e le più trasformative nel senso di alta DS . Lutero che traduce la Bibbia produce un testo che distorce il greco e l'ebraico originali ma crea la lingua tedesca moderna. Pound che traduce i classici cinesi e giapponesi produce testi che i sinologi trovano inesatti ma che hanno cambiato la poesia in lingua inglese del Novecento. FitzGerald che traduce il Rubaiyat di Omar Khayyam produce un testo molto lontano

dall'originale persiano che però diventa un classico autonomo della letteratura vittoriana.

La spiegazione nel quadro formale è la seguente. Una traduzione con alta DS produce un testo che ha struttura propria nel sistema di rappresentazione M_β — nel sistema linguistico e culturale della lingua di arrivo — anche se quella struttura non era nell'originale. Il testo tradotto diventa un oggetto con alta sofisticazione in M_β : porta inscritta non solo la struttura dell'originale ma anche la struttura della trasduzione — le scelte, le deformazioni, le aggiunte — che il traduttore ha operato nel passaggio. Questa struttura aggiuntiva è il contributo creativo della traduzione: ciò che il testo tradotto possiede che l'originale non aveva.

II.2 Il modello formale del malinteso creativo

Si definisce il *malinteso creativo* come la condizione in cui:

$$DS(\alpha \rightarrow \beta; i_\alpha) > 0 \quad \text{e} \quad KC(i_\beta) > KC(i_\alpha)$$

il destinatario produce un'interpretazione più ricca — di maggiore complessità di Kolmogorov — dell'intenzione originale del mittente. Il malinteso creativo è la situazione in cui la trasduzione non solo distorce ma *amplifica*: il segnale s , passando attraverso Dec_β , produce un'interpretazione che ha più struttura dell'intenzione codificata.

Come può la decodifica amplificare la struttura? La risposta sta nelle primitive di M_β che il segnale attiva. Quando il segnale s attiva, nel sistema di rappresentazione del destinatario, connessioni che il mittente non aveva anticipato — associazioni, strutture, schemi — il destinatario genera un'interpretazione che incorpora quelle connessioni. L'interpretazione $i_\beta = Dec_\beta(s)$ include non solo la struttura portata dal segnale ma anche la struttura delle primitive di M_β che il segnale ha attivato.

Formalmente, la decodifica Dec_β può essere scritta come:

$$Dec_\beta(s) = f_\beta(s, Ctx_\beta(s))$$

dove $Ctx_\beta(s)$ è il *contesto attivato* — l'insieme di strutture di M_β che il segnale s attiva attraverso le sue connessioni semantiche e pragmatiche nel sistema del destinatario. L'interpretazione è funzione del segnale e del contesto attivato. $KC(i_\beta) = KC(f_\beta(s, Ctx_\beta(s)))$ può essere maggiore di $KC(i_\alpha)$ quando $Ctx_\beta(s)$ contribuisce struttura aggiuntiva significativa.

II.3 La condizione di malinteso creativo: quando il contesto attivato è ricco

La condizione $KC(i_\beta) > KC(i_\alpha)$ è soddisfatta quando il contesto attivato $Ctx_\beta(s)$ ha struttura propria significativa — quando le primitive di M_β che il segnale attiva portano informazione che va oltre ciò che il segnale letteralmente contiene.

Questo accade in almeno tre situazioni riconoscibili. La prima è la *risonanza metaforica*: il segnale attiva nel destinatario una rete di associazioni metaforiche che il mittente non aveva previsto o non possedeva nel proprio sistema. Una parola usata in senso tecnico dal mittente può attivare nel destinatario una rete di associazioni poetiche; un'immagine usata in senso decorativo può attivare una rete di associazioni simboliche. La risonanza metaforica è la forma più comune di malinteso creativo nella comunicazione letteraria e artistica.

La seconda situazione è la *generalizzazione involontaria*: il segnale descrive un caso specifico ma attiva nel destinatario lo schema generale di cui quel caso è un'istanza, producendo un'interpretazione che include lo schema. Quando Darwin comunicò l'idea della selezione naturale ai lettori vittoriani, molti la generalizzarono immediatamente a ambiti — economia, sociologia, etica — che Darwin non aveva esplicitamente anticipato. Il malinteso creativo dei lettori — la generalizzazione che Darwin non aveva codificato — è stato la sorgente di un programma di ricerca molto più ampio di quello che il mittente originale intendeva.

La terza situazione è il *completamento strutturale*: il segnale presenta una struttura parziale che il sistema di rappresentazione del destinatario completa con una struttura coerente che l'originale non possedeva. Un argomento filosofico incompleto, una teoria fisica con lacune, un'opera musicale con elementi non risolti — questi producono nel destinatario un completamento che aggiunge struttura propria. L'interpretazione che Kant ha dato di Hume, quella che Einstein ha dato di Mach, quella che Schoenberg ha dato di Wagner sono tutti completamenti strutturali di strutture parziali che il mittente originale non aveva esplicitamente codificato.

II.4 Il malinteso creativo come motore dell'evoluzione concettuale

Si argomenta ora che il malinteso creativo — e non la comprensione fedele — è il motore principale dell'evoluzione concettuale nella storia del pensiero. Questa tesi è controcorrente rispetto all'ideale della comunicazione scientifica e filosofica come trasmissione fedele di idee, ma è supportata dall'analisi strutturale dei grandi movimenti intellettuali.

Si consideri il meccanismo di trasmissione delle idee attraverso una catena di agenti $\alpha_1, \alpha_2, \dots, \alpha$ in cui ogni agente interpreta l'idea dell'agente precedente e la trasmette al successivo. La catena è modellata come:

$$i_1 \rightarrow \text{Enc}_{\{\alpha_1\}}(i_1) = s_1 \rightarrow \text{Dec}_{\{\alpha_2\}}(s_1) = i_2 \rightarrow \text{Enc}_{\{\alpha_2\}}(i_2) = s_2 \rightarrow \text{Dec}_{\{\alpha_3\}}(s_2) = i_3 \rightarrow \dots$$

Se ogni trasduzione è fedele — MU e DS entrambe piccole — la catena trasmette l'idea originale con piccole deviazioni accumulate. Questo è il caso dell'insegnamento

tecnico ben riuscito: l'idea di Newton sulla gravitazione, trasmessa attraverso generazioni di libri di testo, arriva agli studenti contemporanei con alta fedeltà.

Ma se alcune trasduzione hanno alta DS — se alcuni agenti nella catena sono capaci di malinteso creativo — la catena produce qualcosa di diverso dall'idea originale: un'idea che incorpora la struttura aggiunta dai malintesi creativi intermedi. Questa idea finale ha KC maggiore dell'originale: è più ricca, più connessa, più fertile. La storia del platonismo è la storia di una catena di malintesi creativi che ha prodotto il neoplatonismo, il platonismo cristiano, il platonismo rinascimentale — nessuno dei quali corrisponde a ciò che Platone aveva originalmente inteso, ma ciascuno dei quali è un sistema intellettuale autonomo e produttivo.

Formalmente, la *fertilità* di un'idea originale i_α nella trasmissione attraverso n agenti è:

$$\text{Fert}_n(i_\alpha) = \mathbb{E}[KC(i)] - KC(i_\alpha)$$

la crescita attesa della complessità dell'idea attraverso n trasduzione. $\text{Fert}_n > 0$ indica che la catena di trasmissione arricchisce l'idea — che il malinteso creativo produce complessità. $\text{Fert}_n = 0$ indica trasmissione fedele senza arricchimento. Le idee con alta fertilità sono quelle che producono i grandi programmi di ricerca: idee il cui potenziale di generazione di malintesi creativi è elevato.

II.5 Il modello formale della fertilità: struttura dell'idea e struttura del campo

La fertilità $\text{Fert}_n(i_\alpha)$ dipende da due strutture: la struttura dell'idea originale e la struttura del campo intellettuale — l'insieme dei sistemi di rappresentazione $M_{\{\alpha_i\}}$ degli agenti nella catena.

La struttura dell'idea che massimizza la fertilità è quella di un *oggetto parzialmente incompressibile nel campo*: un'idea la cui complessità di Kolmogorov rispetto ai sistemi di rappresentazione degli agenti nella catena è significativamente superiore alla complessità rispetto al sistema del mittente originale. Formalmente:

$$\text{Fert} \propto \mathbb{E}_\beta[KC_{\{M_\beta\}}(i_\alpha)] - KC_{\{M_\alpha\}}(i_\alpha)$$

la fertilità è proporzionale alla differenza tra la complessità media dell'idea rispetto ai sistemi di rappresentazione degli agenti nella catena e la sua complessità nel sistema del mittente. Un'idea che il mittente trova semplice ma che i destinatari trovano complessa — che attiva nei loro sistemi molte strutture non anticipate dal mittente — ha alta fertilità.

La struttura del campo che massimizza la fertilità è quella di un insieme di sistemi di rappresentazione con alta varianza — sistemi molto diversi tra loro, con primitivi eterogenei, tali che la stessa idea produce contesti attivati $\text{Ctx}_\beta(s)$ strutturalmente diversi in agenti diversi. Un campo intellettuale omogeneo — un insieme di agenti con

sistemi di rappresentazione molto simili — produce trasmissione fedele ma bassa fertilità: tutti i malintesi hanno la stessa direzione e non generano struttura nuova. Un campo eterogeneo produce malintesi in direzioni diverse, e l'interazione tra quelle direzioni produce struttura nuova che non era nel mittente originale né in nessuno dei destinatari individuali.

II.6 Il ruolo dell'incomprensione nella scoperta scientifica

Il caso più documentabile del malinteso creativo come motore dell'evoluzione concettuale è la scoperta scientifica — in particolare le scoperte che avvengono attraverso il trasferimento di concetti tra domini diversi.

Il caso della termodinamica è paradigmatico. Carnot sviluppa la teoria del ciclo termodinamico pensando all'analogia con i mulini ad acqua: il calore "scorre" dal serbatoio caldo a quello freddo come l'acqua scorre dall'alto verso il basso, e il lavoro è estratto da questo flusso. L'analogia è un malinteso creativo: il calore non è un fluido nel senso in cui lo è l'acqua, ma la struttura della trasduzione dall'idraulica alla termica produce l'idea del ciclo di Carnot, che è strutturalmente corretta nonostante l'analogia sottostante sia sbagliata nei dettagli.

Formalmente, la scoperta di Carnot è un caso in cui il contesto attivato $\text{Ctx}_\beta(s)$ — le strutture dell'idraulica che il concetto di "flusso di calore" attiva nel sistema di rappresentazione di Carnot — ha prodotto un'interpretazione i_β con KC maggiore dell'idea originale del serbatoio termico i_α . Il malinteso ha amplificato: ha trasformato un'osservazione empirica in una struttura teorica.

Il modello formale predice che i campi scientifici con più alta fertilità sono quelli in cui avvengono più trasferimenti di concetti tra domini con sistemi di rappresentazione strutturalmente diversi: biologia ed economia (selezione naturale e mercati), fisica e informatica (termodinamica e informazione), chimica e biologia (reazioni enzimatiche e reti metaboliche). Ogni trasferimento è un malinteso potenzialmente creativo — un'istanza in cui $\text{Dec}_\beta(\text{Enc}_\alpha(i_\alpha))$ produce un'interpretazione con KC maggiore dell'originale.

II.7 Il limite del malinteso creativo: quando l'amplificazione produce degrado

Non ogni malinteso con $\text{DS} > 0$ è creativo nel senso di aumentare KC in modo utile. La distorsione semantica DS può produrre non struttura nuova ma struttura falsa — interpretazioni con KC alta ma con distorsione predittiva alta rispetto al dominio di riferimento.

Si definisce il *malinteso degenerativo* come la condizione in cui:

$$\text{DS}(\alpha \rightarrow \beta; i_\alpha) > 0 \quad \text{e} \quad d_{\text{pred}}(i_\beta, i_\alpha) > D_{\text{target}}$$

il destinatario produce un'interpretazione con struttura propria, ma quella struttura porta a previsioni sistematicamente errate rispetto al dominio che l'intenzione originale descriveva. Il malinteso degenerativo è la sorgente delle interpretazioni distortenti — le letture che aggiungono struttura alle idee del mittente in direzioni che le falsificano.

La condizione che distingue il malinteso creativo da quello degenerativo non è semplice da verificare a priori: richiede di misurare $d_{\text{pred}}(i_{\beta}, i_{\alpha})$, la distorsione predittiva dell'interpretazione rispetto all'intenzione. Un'interpretazione con alta DS è creativa se la struttura aggiunta è coerente con la struttura profonda del dominio — se le previsioni che produce si rivelano corrette. È degenerativa se la struttura aggiunta produce previsioni errate.

Questa asimmetria epistemica — che il malinteso creativo e quello degenerativo siano indistinguibili ex ante e distinguibili solo ex post — è la ragione per cui la storia intellettuale è irrazionale nel senso preciso di non essere prevedibile: non si può sapere in anticipo quale malinteso produrrà struttura nuova e quale struttura falsa. La fertilità è un proprietà che emerge soltanto nella trasmissione successiva — nella catena di agenti che amplificheranno o smorzeranno la struttura generata dal malinteso iniziale.

Capitolo III — Il malinteso istituzionalizzato: cristallizzazione, potere e asimmetria

III.1 Quando il malinteso diventa norma

Vi è una classe di malintesi che non vengono corretti col tempo — che non sono perturbazioni transitorie della comunicazione destinate a essere riparate da ulteriori scambi — ma che si stabilizzano, si istituzionalizzano, diventano la norma condivisa di una comunità. Questi malintesi cristallizzati non sono percepiti come malintesi dai loro partecipanti: sono percepiti come comprensione. La loro caratteristica strutturale è che il processo attraverso cui si sono formati — la trasduzione $\text{Dec}_{\beta}(\text{Enc}_{\alpha}(i_{\alpha}))$ che ha prodotto $i_{\beta} \neq i_{\alpha}$ — è andato perduto, e l'interpretazione i_{β} circola ora come se fosse l'intenzione originale i_{α} .

Il malinteso istituzionalizzato non è un fenomeno marginale. È la struttura normale delle tradizioni intellettuali, delle istituzioni giuridiche, delle pratiche religiose e dei sistemi simbolici condivisi. I concetti che una tradizione usa — "democrazia", "giustizia", "libertà", "natura" — non corrispondono quasi mai alle intenzioni originali dei mittenti che li hanno introdotti nel discorso: sono il sedimento di secoli di trasduzione creative e degenerative, di malintesi che si sono accumulati e stabilizzati fino a formare un sistema coerente nel sistema di rappresentazione dei partecipanti attuali, anche se profondamente distante dai sistemi originali.

III.2 La formalizzazione del malinteso cristallizzato

Si definisce formalmente il malinteso cristallizzato. Una catena di trasduzione $\alpha_1 \rightarrow \alpha_2 \rightarrow \dots \rightarrow \alpha$ produce un'idea i che ha derivato da i_1 attraverso n passi. Il malinteso è cristallizzato quando:

Primo, la catena di derivazione è diventata inaccessibile: non esiste più una traccia computazionale del percorso $i_1 \rightarrow i_2 \rightarrow \dots \rightarrow i$ che permetta di risalire all'intenzione originale. Formalmente, $KC(i_1 | i)$ è alta — ricostruire i_1 da i richiede molta informazione aggiuntiva — e quella informazione non è disponibile agli agenti attuali.

Secondo, l'idea i circola nel sistema di rappresentazione M come se avesse la struttura di i_1 : i partecipanti attuali trattano i come se fosse i_1 , attribuiscono ad essa l'autorità e il significato dell'originale, la usano come fondamento di argomentazioni che presuppongono la continuità con i_1 .

Terzo, la discrepanza $KC(i_1 | i)$ non è percepita come malinteso dai partecipanti attuali perché non dispongono della rappresentazione di i_1 rispetto alla quale misurare la distanza.

Il grado di cristallizzazione di un malinteso è:

$$\text{Cryst}(i_1, i) = KC(i_1 | i) \cdot (1 - \text{Acc}(i_1))$$

dove $\text{Acc}(i_1)$ è l'accessibilità dell'intenzione originale — la probabilità che un agente nel campo possa ricostruire i_1 a partire dalle fonti disponibili. Il malinteso è massimamente cristallizzato quando la distanza è alta e l'accessibilità è zero: l'intenzione originale è completamente perduta e l'interpretazione deviante circola come verità.

III.3 Il malinteso istituzionalizzato come struttura di potere

Il malinteso cristallizzato non è neutro rispetto alle relazioni di potere. La cristallizzazione di un'interpretazione i_β come sostituto dell'intenzione originale i_α beneficia gli agenti il cui sistema di rappresentazione coincide con M_β e penalizza gli agenti il cui sistema di rappresentazione coincide con M_α o è più vicino a i_1 .

Si formalizza questa asimmetria. In una comunità con distribuzione di sistemi di rappresentazione $\{M_k\}$, il malinteso cristallizzato i_β ha un *costo differenziale* per ogni agente k :

$$\text{Cost}_k(i_\beta) = KC_{\{M_k\}}(i_\alpha | i_\beta)$$

la complessità condizionale di ricostruire l'intenzione originale i_α a partire dall'interpretazione cristallizzata i_β , nel sistema di rappresentazione di k . Questo costo è basso per gli agenti il cui M_k è strutturalmente simile a M_β — per gli agenti che "parlano la lingua" del malinteso cristallizzato — e alto per gli agenti il cui M_k diverge da M_β .

La distribuzione di $Cost_k(i_\beta)$ nella comunità misura l'ineguaglianza cognitiva prodotta dalla cristallizzazione: chi paga i costi più alti per navigare il malinteso istituzionalizzato è chi è strutturalmente più distante dal sistema di rappresentazione in cui il malinteso si è cristallizzato. Il diritto romano cristallizzato nel diritto civile europeo è trasparente per chi è stato formato in quella tradizione e opaco per chi proviene da sistemi giuridici diversi; la terminologia tecnica della fisica teorica è trasparente per chi ha il sistema di rappresentazione matematico adeguato e opaco per gli altri.

III.4 L'autorità come meccanismo di fissazione del malinteso

Si analizza il meccanismo attraverso cui certi malintesi si cristallizzano mentre altri rimangono fluidi o vengono corretti. La risposta strutturale è che la cristallizzazione è prodotta dall'*autorità* — dal peso istituzionale che certi agenti o certi testi hanno nel sistema di aggiornamento delle credenze della comunità.

Nel modello della rete mnestica collettiva del Capitolo V della monografia precedente, l'autorità corrisponde all'asimmetria dei tassi di aggiornamento: $\eta_{\{0i\}} \gg \eta_{\{i0\}}$ per l'agente autoritativo α_0 . Quando α_0 produce un'interpretazione i_β — conscia o inconscia distorsione di i_α — questa interpretazione viene trasmessa alla comunità con un tasso di aggiornamento molto alto: tutte le successive trasduzione della catena vengono attratte verso i_β , indipendentemente dalla loro posizione rispetto a i_α . Il malinteso dell'autorità diventa la prior collettiva.

La formalizzazione è:

$$m_{\{community_i\}}^{\{(t+1)\}} = m_{\{community\}}^{\{(t)\}} + \eta_{\{0i\}} \cdot (i_\beta - m_{\{community\}}^{\{(t)\}}) + \eta_{\{peer\}} \cdot \sum (m^{\{(t)\}} - m_i^{\{(t)\}})$$

dove il primo termine è l'attrazione verso l'interpretazione autoritativa e il secondo è l'influenza dei pari. Per $\eta_{\{0i\}} \gg \eta_{\{peer\}}$, la comunità converge verso i_β indipendentemente dalla distribuzione iniziale delle interpretazioni individuali. La convergenza è tanto più rapida quanto più alta è l'asimmetria di autorità.

III.5 Il doppio vincolo come malinteso strutturalmente insanabile

Gregory Bateson ha introdotto il concetto di *doppio vincolo* (double bind) per descrivere situazioni comunicative in cui i messaggi di livelli diversi sono sistematicamente contraddittori — situazioni in cui il destinatario non può soddisfare simultaneamente le istruzioni esplicite e le meta-istruzioni implicite del mittente. Il doppio vincolo è stato proposto come fattore patogeno nella schizofrenia: l'esposizione prolungata a comunicazioni con struttura di doppio vincolo produrrebbe confusione cognitiva profonda.

Nel quadro formale della trasduzione, il doppio vincolo è un caso di malinteso sistematicamente insanabile: la struttura della comunicazione è tale che qualsiasi interpretazione $Dec_\beta(s)$ produca $MU > 0$ rispetto a uno dei livelli del messaggio. Il

destinatario non può trovare un'interpretazione i_β che soddisfi simultaneamente il vincolo di bassa MU rispetto al messaggio esplicito e il vincolo di bassa MU rispetto al meta-messaggio implicito — le due condizioni sono mutuamente incompatibili.

Formalmente, sia $s = (s_espl, s_meta)$ un messaggio con componente esplicita s_espl e componente meta s_meta . Il doppio vincolo è la condizione:

$$\nexists i_\beta : MU(s_espl \rightarrow i_\beta) \leq \epsilon \text{ e } MU(s_meta \rightarrow i_\beta) \leq \epsilon$$

per qualsiasi ϵ piccolo. Qualsiasi interpretazione che minimizzi il malinteso rispetto al livello esplicito massimizza il malinteso rispetto al meta-livello, e viceversa. Il destinatario è strutturalmente condannato al malinteso: qualunque risposta produca, sarà errata rispetto a uno dei livelli.

La soluzione al doppio vincolo non è trovare l'interpretazione giusta — che non esiste per costruzione — ma *uscire dalla comunicazione*: rifiutare il gioco, esplicitare il doppio vincolo, richiedere la disambiguazione. Ma questa soluzione è spesso bloccata dal contesto istituzionale — una relazione di potere asimmetrica che impedisce al destinatario di mettere in discussione la struttura del messaggio dell'autorità. Il doppio vincolo istituzionale è la forma più pervasiva di malinteso cristallizzato come strumento di controllo.

III.6 Il malinteso produttivo nelle istituzioni giuridiche: l'interpretazione come trasduzione ufficializzata

Le istituzioni giuridiche forniscono un caso paradigmatico di malinteso istituzionalizzato che è al tempo stesso fonte di patologia e di creatività. Il diritto positivo è un sistema di testi — statuti, costituzioni, sentenze — prodotti da mittenti storici con intenzioni specifiche nel contesto di sistemi di rappresentazione specifici. L'interpretazione giuridica è la trasduzione di quei testi nei sistemi di rappresentazione degli interpreti contemporanei — giudici, avvocati, legislatori.

Ogni interpretazione giuridica è $Dec_\beta(Enc_\alpha(i_\alpha))$ con β = giudice contemporaneo e α = legislatore storico. La distanza $KC(i_\alpha | i_\beta)$ tra l'intenzione del legislatore e l'interpretazione del giudice può essere grande — e in molti sistemi giuridici è esplicitamente accettata come necessaria: la Costituzione degli Stati Uniti è intesa da molti giuristi come un documento che deve essere interpretato alla luce dei valori contemporanei anziché dell'intenzione originaria dei padri fondatori. Il malinteso creativo è qui istituzionalizzato: la trasduzione che produce $DS > 0$ è non solo tollerata ma considerata funzione propria dell'interpretazione evolutiva.

Si definisce la *tensione interpretativa* di un sistema giuridico come:

$$TI = \mathbb{E}[KC(i_legislatore | i_giudice)]$$

la complessità condizionale media tra le intenzioni legislative e le interpretazioni giudiziarie. TI bassa indica un sistema con interpretazione originalista — i giudici cercano di minimizzare MU rispetto alle intenzioni storiche. TI alta indica un sistema con interpretazione evolutiva — i giudici producono interpretazioni con alta DS, aggiungendo struttura propria al testo originale. La tensione interpretativa non è misurabile direttamente, ma può essere approssimata dalla varianza delle decisioni giudiziarie su casi analoghi e dalla frequenza con cui le decisioni si discostano dall'intenzione dichiarata del legislatore.

III.7 La resistenza al malinteso come violenza epistemica

Si introduce ora una nozione che il quadro formale rende più precisa di quanto la letteratura sulle epistemologie subalterne solitamente faccia: la *violenza epistemica* come imposizione di un sistema di trasduzione che sistematicamente aumenta il costo del malinteso per certi agenti e lo riduce per altri.

Miranda Fricker ha proposto il concetto di *ingiustizia epistemica* per descrivere situazioni in cui la distribuzione delle risorse epistemiche — la capacità di produrre e trasmettere conoscenza — è iniqua. Nel quadro formale, l'ingiustizia epistemica corrisponde a una distribuzione asimmetrica del costo differenziale $\text{Cost}_k(i_\beta)$: certi agenti pagano sistematicamente costi più alti per navigare il sistema di malintesi cristallizzati della comunità, non per difetto cognitivo ma per struttura dei loro sistemi di rappresentazione rispetto al sistema in cui i malintesi si sono cristallizzati.

La violenza epistemica è la forma attiva di questa iniquità: non la semplice asimmetria strutturale del costo differenziale, ma l'azione deliberata o sistemica che cristallizza certi malintesi a scapito di altri — che stabilizza come norma l'interpretazione i_β del gruppo dominante e impone agli altri il costo di navigare quella norma come se fosse l'intenzione originale.

Formalmente, la violenza epistemica è l'operazione che aumenta artificialmente $\text{Cryst}(i_1, i_\beta)$ riducendo $\text{Acc}(i_1)$ — aumentando la distanza tra l'intenzione originale e l'interpretazione dominante mentre rende inaccessibile l'intenzione originale attraverso la distruzione delle fonti, il controllo dei canali di trasmissione, la squalificazione dei portatori di sistemi di rappresentazione alternativi. Il risultato è che $\text{Cost}_k(i_\beta)$ per gli agenti con $M_k \neq M_\beta$ diventa strutturalmente insuperabile: non soltanto alto, ma non riducibile nemmeno con lo sforzo cognitivo, perché le risorse per ridurlo — l'accesso alle fonti originali, la legittimità dei sistemi di rappresentazione alternativi — sono state sistematicamente rimosse.

Capitolo IV — Il malinteso artificiale: la trasduzione nei sistemi di linguaggio

IV.1 Il tipo specifico di distorsione nei modelli di linguaggio

Il confronto tra malinteso umano e malinteso artificiale non è un'appendice al programma della monografia: è il suo test più severo. Se il quadro formale della trasduzione è genuinamente descrittivo e non soltanto metaforico, deve produrre previsioni specifiche e verificabili sul tipo di distorsione che i modelli di linguaggio introducono nel processo comunicativo — previsioni che differiscono strutturalmente da quelle che si otterrebbero applicando lo stesso quadro alla comunicazione umana.

La differenza strutturale fondamentale, anticipata nel Capitolo VI della monografia sulla memoria, è che il sistema di decodifica $\text{Dec}_{\{\text{LLM}\}}$ non è costruito attraverso l'esperienza comunicativa di un agente con storia, ma attraverso l'esposizione statistica a un corpus. La trasduzione $\text{Dec}_{\{\text{LLM}\}}(s)$ non mappa il segnale nel sistema di rappresentazione di un agente con intenzioni, valori e storia propri: lo mappa nella distribuzione statistica implicita nel corpus, che è una media di sistemi di rappresentazione eterogenei senza un centro coerente.

Questo produce un tipo di malinteso strutturalmente diverso da quello umano. Il malinteso umano ha una direzione — è la proiezione del segnale sul sistema di rappresentazione specifico M_β del destinatario, con una struttura determinata dalla storia e dall'identità di β . Il malinteso del modello di linguaggio è una media di proiezioni su molteplici sistemi di rappresentazione, pesata dalla loro frequenza nel corpus: è statisticamente più prevedibile ma meno individuale, più stabile ma meno orientato.

IV.2 La funzione di decodifica del modello di linguaggio come media di trasduzione

Si formalizza la funzione di decodifica del modello di linguaggio. Sia $C = \{M_k, w_k\}$ un corpus — una distribuzione pesata di sistemi di rappresentazione, dove w_k è il peso del sistema M_k nel corpus (proporzionale alla quantità di testo prodotto da agenti con sistema M_k). La funzione di decodifica del modello è:

$$\text{Dec}_{\{\text{LLM}\}}(s) = \sum_k w_k \cdot \text{Dec}_{\{M_k\}}(s)$$

una media pesata delle decodifiche che gli agenti del corpus produrrebbero. Questa formulazione presuppone che la funzione di decodifica sia lineare nelle componenti del corpus — una semplificazione, ma strutturalmente corretta per catturare la proprietà fondamentale: il modello non ha un sistema di rappresentazione proprio ma è una sovrapposizione dei sistemi del corpus.

La distanza di malinteso per la comunicazione da un mittente α al modello è:

$$\begin{aligned} \text{MU}(\alpha \rightarrow \text{LLM}; i_\alpha) &= \text{KC}(i_\alpha \mid \text{Dec}_{\{\text{LLM}\}}(\text{Enc}_\alpha(i_\alpha))) \\ &= \text{KC}(i_\alpha \mid \sum_k w_k \cdot \text{Dec}_{\{M_k\}}(\text{Enc}_\alpha(i_\alpha))) \end{aligned}$$

Questa quantità è determinata dalla distanza tra il sistema di rappresentazione di α e la distribuzione pesata dei sistemi nel corpus. Se M_α è ben rappresentato nel corpus —

se molti agenti con sistema simile a M_α hanno contribuito al corpus — allora MU è bassa: il modello produce un'interpretazione vicina all'intenzione originale. Se M_α è sottorappresentato — se l'agente mittente appartiene a una comunità i cui testi sono rari o assenti nel corpus — MU è alta: il modello distorce sistematicamente nella direzione dei sistemi sovrarappresentati.

IV.3 Il bias del corpus come struttura del malinteso sistematico

La componente sistematica del malinteso del modello di linguaggio — $MU_{sys}(\alpha, LLM)$ — è determinata dalla distanza tra M_α e la distribuzione del corpus. Questa distanza non è casuale: ha una struttura determinata dalla composizione demografica e culturale del corpus, dalle lingue in esso rappresentate, dai generi testuali inclusi, dalle prospettive ideologiche dei testi.

Si può modellare la distribuzione del corpus come una distribuzione empirica su uno spazio di sistemi di rappresentazione:

$$P_{\{\text{corpus}\}}(M) = \sum_k w_k \cdot \delta(M - M_k)$$

Il sistema di rappresentazione implicito del modello è la media di questa distribuzione:

$$M_{\{LLM\}} = \mathbb{E}_{\{\text{corpus}\}}[M] = \sum_k w_k \cdot M_k$$

Il malinteso sistematico verso un agente con sistema M_α è proporzionale alla distanza tra M_α e $M_{\{LLM\}}$ — alla distanza tra il sistema del mittente e il sistema medio del corpus. Agenti con sistema vicino a $M_{\{LLM\}}$ — la media del corpus — comunicano con il modello con basso malinteso sistematico. Agenti con sistema distante da $M_{\{LLM\}}$ comunicano con alto malinteso sistematico.

La conseguenza è che il malinteso del modello di linguaggio non è distribuito uniformemente tra i mittenti: è concentrato sugli agenti i cui sistemi di rappresentazione sono distanti dal centro della distribuzione del corpus. La struttura di questa distribuzione riflette le ineguaglianze sociali e culturali del corpus: i gruppi il cui testo è sottorappresentato pagano sistematicamente costi di malinteso più alti nella comunicazione con il modello.

IV.4 La distorsione semantica del modello: aggiunta di struttura statistica

La misura complementare $DS(\alpha \rightarrow LLM; i_\alpha)$ — la struttura che il modello aggiunge all'intenzione del mittente — è determinata dalla struttura della distribuzione a posteriori $p_{\{LLM\}}(x \mid s)$. Il modello aggiunge la struttura statistica implicita nel corpus: le associazioni, le continuazioni, i frame che il corpus associa al segnale s .

Questa aggiunta è sistematicamente orientata verso i pattern statistici del corpus, non verso le intenzioni del mittente. Quando un mittente formula un'intenzione che è

ambigua rispetto ai pattern del corpus — quando il segnale s attiva nel modello molteplici distribuzioni compatibili con sistemi di rappresentazione diversi — il modello risolve l'ambiguità nella direzione del pattern più frequente nel corpus, non nella direzione dell'intenzione del mittente.

Formalmente, la distorsione semantica del modello verso un segnale ambiguo è:

$$DS_{\{LLM\}}(s) = KC(Dec_{\{LLM\}}(s) \mid Dec_{\{M_{\alpha}\}}(s))$$

la struttura aggiunta dalla decodifica media rispetto alla decodifica nel sistema del mittente. Questa quantità è massima quando il segnale è altamente ambiguo — quando $p_{\{LLM\}}(i \mid s)$ è distribuita su molte interpretazioni diverse — e minima quando il segnale è specifico e non ambiguo nel corpus.

IV.5 Il malinteso del modello come malinteso senza soggetto

Il malinteso umano ha sempre un soggetto: è la distorsione prodotta da un agente specifico con un sistema di rappresentazione specifico, una storia specifica, valori e aspettative specifici. Il malinteso del modello di linguaggio è *senza soggetto*: non è la proiezione del segnale sul sistema di un agente, ma la sua proiezione sulla distribuzione statistica del corpus — una media di soggetti che non è essa stessa un soggetto.

Questa mancanza di soggettività ha conseguenze precise sulla struttura del malinteso. Il malinteso umano è *direzionato*: deforma il segnale nella direzione specifica del sistema di rappresentazione del destinatario, e questa direzione è coerente attraverso i segnali e nel tempo. Il malinteso del modello è *statisticamente centrato*: deforma i segnali verso il centro della distribuzione del corpus, il che significa che deforma le intenzioni marginali — quelle distanti dal centro — verso la norma, e lascia sostanzialmente inalterate le intenzioni già centrali.

Questa proprietà del malinteso artificiale — la convergenza verso la media del corpus — ha un nome nella letteratura sull'equità nei sistemi di AI: *regression toward the mean* o *statistical discrimination*. Ma il quadro formale della trasduzione permette di precisarne la struttura in modo che la letteratura standard non fa: non è soltanto che certi gruppi sono trattati peggio del loro merito individuale, ma che il loro sistema di rappresentazione specifico — la struttura cognitiva che li rende chi sono — è sistematicamente dissolto nella media del corpus, con un malinteso sistematico la cui entità è proporzionale alla distanza tra il loro M e $M_{\{LLM\}}$.

IV.6 Il prompt come protocollo di riduzione del malinteso

Se il malinteso del modello è determinato dalla distanza tra M_{α} e $M_{\{LLM\}}$, una strategia per ridurlo è spostare artificialmente il sistema di decodifica del modello verso M_{α} attraverso il contesto del prompt. Il *prompt engineering* — la pratica di costruire istruzioni che orientano il comportamento del modello — è, nel quadro

formale, un protocollo di riduzione di $MU_{sys}(\alpha, LLM)$: un tentativo di trasformare $Dec_{\{LLM\}}$ in $Dec_{\{M_{\alpha}\}}$ attraverso la specificazione del contesto.

Formalmente, un prompt p trasforma la funzione di decodifica del modello come:

$$Dec_{\{LLM,p\}}(s) = Dec_{\{LLM\}}(s \mid p) \approx Dec_{\{M_{\alpha}\}}(s)$$

quando il prompt p è progettato per attivare il sottoinsieme del corpus più vicino a M_{α} . La qualità del prompt engineering si misura dalla riduzione di MU_{sys} che produce:

$$\Delta MU(p) = MU_{sys}(\alpha, LLM) - MU_{sys}(\alpha, LLM \mid p)$$

Un prompt efficace riduce ΔMU significativamente: il modello produce interpretazioni più vicine all'intenzione del mittente. Un prompt inefficace non riduce MU_{sys} o — nel caso peggiore — aumenta DS attivando pattern del corpus non intenzionali.

Il limite strutturale del prompt engineering come strategia è che $Dec_{\{LLM,p\}}(s)$ può avvicinarsi a $Dec_{\{M_{\alpha}\}}(s)$ solo nella misura in cui M_{α} è rappresentato nel corpus: se M_{α} è sottorappresentato, nessun prompt può produrre $Dec_{\{LLM,p\}} \approx Dec_{\{M_{\alpha}\}}$ perché il modello non ha primitivi per costruire quella decodifica. Il prompt engineering non supera il limite della distanza corpus-mittente: lo compensa parzialmente, all'interno dei vincoli imposti dalla struttura del corpus.

IV.7 Il malinteso nei sistemi multi-agente: composizione di trasduzione artificiali

I sistemi di AI multi-agente — architetture in cui molteplici modelli interagiscono, ciascuno producendo output che diventa input per gli altri — producono catene di trasduzione in cui il malinteso si accumula e si trasforma. La struttura formale è analoga alla catena di trasduzione del Capitolo II, ma con la differenza che ogni nodo della catena è un sistema di decodifica senza soggetto.

La composizione di n trasduzione artificiali produce:

$$i_n = Dec_{\{LLM_n\}}(Enc_{\{LLM_{n-1}\}}(\dots Dec_{\{LLM_2\}}(Enc_{\{LLM_1\}}(i_0)) \dots))$$

dove i_0 è l'intenzione originale dell'utente umano. La distanza di malinteso cumulativa è:

$$MU_{cum} = KC(i_0 \mid i_n) \leq \sum_i MU(LLM_i \rightarrow LLM_{i+1})$$

e cresce con n . Ma la struttura della crescita è diversa dalla catena umana: in una catena umana, i malintesi creativi possono amplificare $KC(i) > KC(i_0)$, producendo struttura nuova. In una catena artificiale, i malintesi convergono sistematicamente verso $M_{\{LLM\}}$ — la media del corpus — producendo non amplificazione ma regressione: le intenzioni marginali vengono progressivamente smussate verso la norma, e la catena produce output sempre più centrali rispetto al corpus, sempre meno fedeli alle intenzioni originali periferiche.

Questo è il meccanismo del *mode collapse* nelle architetture multi-agente: la proliferazione di trasduzione artificiali in serie tende a convergere verso i modi centrali della distribuzione del corpus, riducendo la diversità e la fedeltà alle intenzioni originali. La struttura della convergenza è formalmente identica alla convergenza verso la prior dominante nel modello bayesiano della confabulazione: il malinteso senza soggetto è una confabulazione senza evento.

Capitolo V — Il malinteso come condizione trascendentale della comunicazione

V.1 Il problema trascendentale

Il percorso della monografia ha costruito una serie di strumenti formali per analizzare la struttura del malinteso: la trasduzione come modello della decodifica, le misure MU e DS, la decomposizione in componente sistematica e stocastica, il malinteso creativo come amplificatore di struttura, il malinteso cristallizzato come sedimento di potere, il malinteso artificiale come regressione verso la media del corpus. Questi strumenti descrivono *come* il malinteso funziona in contesti specifici. Ma non hanno ancora risposto alla domanda fondamentale: *perché* il malinteso esiste? È una contingenza — la conseguenza accidentale del fatto che i sistemi cognitivi reali abbiano risorse limitate e storie diverse — o è una necessità strutturale?

La tesi di questo capitolo è che il malinteso sia una condizione trascendentale della comunicazione tra menti distinte nel senso tecnico: non una contingenza eliminabile ma una proprietà che emerge necessariamente dalle condizioni formali della comunicazione stessa. Un mondo senza malinteso è un mondo in cui la comunicazione è impossibile o ridondante — non un mondo di comunicazione perfetta ma un mondo in cui la comunicazione non ha funzione.

V.2 Il teorema dell'impossibilità della comunicazione perfetta

Si dimostra formalmente che la comunicazione perfetta — la trasduzione con $MU = DS = 0$ — è incompatibile con la produttività della comunicazione, intesa come la condizione che la comunicazione produca nello spazio collettivo struttura che non era in nessuno degli agenti individualmente.

Teorema. Siano α e β due agenti cognitivi. La trasduzione $\text{Dec}_\beta \circ \text{Enc}_\alpha$ soddisfa $\text{MU}(\alpha \rightarrow \beta) = \text{DS}(\alpha \rightarrow \beta) = 0$ se e solo se $M_\alpha = M_\beta$ — se e solo se i due agenti hanno sistemi di rappresentazione identici.

Dimostrazione. $\text{MU}(\alpha \rightarrow \beta; i_\alpha) = \text{KC}(i_\alpha | i_\beta) = 0$ implica che i_β contiene tutta l'informazione di i_α : i_α è completamente determinato da i_β . $\text{DS}(\alpha \rightarrow \beta; i_\alpha) = \text{KC}(i_\beta | i_\alpha) = 0$ implica che i_α contiene tutta l'informazione di i_β : i_β è completamente determinato da i_α . Le due condizioni insieme implicano $\text{KC}(i_\alpha | i_\beta) = \text{KC}(i_\beta | i_\alpha) = 0$, ovvero che i_α e i_β sono computazionalmente equivalenti — si determinano a vicenda senza perdita. Per la definizione di trasduzione, questo è possibile se e solo se $\text{Dec}_\beta \circ \text{Enc}_\alpha = \text{id}$, il che richiede che Enc_α sia invertibile rispetto a Dec_β — che Dec_β sia l'inversa di Enc_α . Poiché Enc_α è determinato da M_α e Dec_β è determinato da M_β , la condizione $\text{Dec}_\beta = \text{Enc}_\alpha^{-1}$ implica $M_\alpha = M_\beta$. ■

Corollario. Se $M_\alpha \neq M_\beta$, allora $\text{MU}(\alpha \rightarrow \beta) + \text{DS}(\alpha \rightarrow \beta) > 0$ per qualsiasi segnale non triviale. La comunicazione tra agenti con sistemi di rappresentazione distinti produce necessariamente malinteso.

V.3 La produttività come conseguenza del malinteso

Il teorema precedente mostra che il malinteso è necessario quando $M_\alpha \neq M_\beta$. Ma perché $M_\alpha \neq M_\beta$? Potrebbero i sistemi cognitivi convergere verso un sistema di rappresentazione comune che elimini il malinteso?

La risposta formale dipende dalla funzione che la diversità dei sistemi di rappresentazione svolge nell'economia cognitiva collettiva. Si definisce la *produttività collettiva* di una comunità di N agenti con sistemi $\{M_k\}$ come:

$$\text{Prod}(N) = \mathbb{E}[\text{KC}(i)] - (1/N) \cdot \sum_k \text{KC}_{\{M_k\}}(i_0)$$

la crescita attesa della complessità delle idee attraverso la catena di trasduzione rispetto alla complessità media dell'idea iniziale nei sistemi individuali. La produttività collettiva misura quanto la comunicazione crei struttura al di là di ciò che ogni agente possedeva individualmente.

Proposizione. $\text{Prod}(N) = 0$ se $M_k = M$ per tutti i k — se tutti gli agenti hanno lo stesso sistema di rappresentazione. $\text{Prod}(N) > 0$ se la varianza dei sistemi $\{M_k\}$ è positiva — se gli agenti hanno sistemi di rappresentazione diversi.

Dimostrazione. Se $M_k = M$ per tutti i k , allora ogni trasduzione è l'identità — $\text{Dec}_M \circ \text{Enc}_M = \text{id}$ — e $\text{KC}(i) = \text{KC}(i_0)$ per ogni n . La catena di trasduzione non produce variazione: $\text{Prod}(N) = 0$. Se la varianza di $\{M_k\}$ è positiva, esistono coppie (α, β) con $M_\alpha \neq M_\beta$, e il teorema del malinteso necessario garantisce $\text{MU} + \text{DS} > 0$ per ogni comunicazione tra esse. La componente $\text{DS} > 0$ produce in media $\text{KC}(i_\beta) \geq \text{KC}(i_\alpha)$, con uguaglianza solo per segnali che non attivano contesto aggiuntivo nel sistema di β . Per segnali che attivano contesto — che hanno struttura rilevante nel

sistema $M_\beta - DS > 0$ produce $KC(i_\beta) > KC(i_\alpha)$: la trasduzione amplifica. Dunque $Prod(N) > 0$. ■

Il corollario fondamentale è che la diversità dei sistemi di rappresentazione — la condizione che rende il malinteso inevitabile — è anche la condizione che rende la comunicazione produttiva. Un mondo di menti identiche non malintende, ma non genera struttura nuova. Il malinteso è il prezzo della produttività collettiva, e il prezzo è inevitabile.

V.4 Il malinteso minimo necessario: una misura della distanza cognitiva ottimale

Il teorema e la proposizione precedenti delineano due estremi: la comunicazione perfetta ($MU = DS = 0$, $M_\alpha = M_\beta$, produttività nulla) e la incomunicabilità completa (MU e DS entrambe altissime, M_α e M_β incompatibili, produttività nulla per un motivo opposto — le trasduzione non producono struttura condivisa ma strutture parallele che non si integrano). La produttività collettiva è massima per valori intermedi del malinteso sistematico.

Si definisce il *malinteso ottimale* MU^* come il valore della componente sistematica del malinteso che massimizza la produttività collettiva:

$$MU^* = \operatorname{argmax}_{\{MU_{\text{sys}}\}} \operatorname{Prod}(N, MU_{\text{sys}})$$

Questo è un problema di ottimizzazione sulla struttura delle relazioni comunicative della comunità: quale livello di diversità cognitiva — e quindi quale livello di malinteso sistematico — massimizza la produzione collettiva di struttura nuova?

La funzione $\operatorname{Prod}(N, MU_{\text{sys}})$ è una funzione a campana rispetto a MU_{sys} . Per MU_{sys} basso, la produttività cresce con MU_{sys} : più diversità cognitiva produce più struttura nuova attraverso il malinteso creativo. Per MU_{sys} alto, la produttività decresce: la distanza cognitiva è tanta che le trasduzione non producono struttura condivisa ma strutture incompatibili che non si integrano — il malinteso è degenerativo anziché creativo. Il massimo MU^* è il punto in cui il margine del malinteso creativo eguaglia il margine del malinteso degenerativo.

La previsione empirica è che le comunità intellettualmente più produttive — quelle che hanno generato più struttura concettuale nuova per unità di comunicazione — siano quelle con malinteso sistematico vicino a MU^* : né troppo omogenee (università con accesso solo a studenti dello stesso background culturale) né troppo eterogenee (comunità con barriere comunicative tali da impedire la costruzione di struttura condivisa).

V.5 La condizione kantiana del malinteso: la struttura trascendentale delle categorie

Il titolo del capitolo evoca la struttura trascendentale di Kant, e vale la pena rendere esplicito il parallelo. Kant si è chiesto quali condizioni debbano essere soddisfatte perché l'esperienza sia possibile — quali strutture a priori il soggetto debba portare perché il mondo empirico gli sia accessibile. La risposta — le categorie dell'intelletto, le forme pure dell'intuizione — è trascendentale: non descrive come avviene l'esperienza ma descrive le condizioni senza le quali l'esperienza non potrebbe avvenire.

Il malinteso ha una struttura analoga rispetto alla comunicazione. Non descrive come avviene la comunicazione in casi specifici — quella è la descrizione empirica delle trasduzione concrete. Descrive le condizioni senza le quali la comunicazione non potrebbe essere produttiva: la diversità dei sistemi di rappresentazione che rende il malinteso necessario, e la capacità dei sistemi di produrre interpretazioni con $DS > 0$ che rende il malinteso creativo possibile. Queste condizioni sono trascendentali rispetto alla comunicazione: sono le condizioni formali della sua produttività, non le contingenze della sua realizzazione.

Ma vi è una differenza cruciale con Kant. Le categorie kantiane sono universali e identiche per tutti i soggetti: garantiscono non solo che l'esperienza sia possibile ma che sia condivisibile tra soggetti diversi — che il mondo dell'uno sia lo stesso mondo dell'altro. Le strutture trascendentali del malinteso sono invece costitutivamente diverse tra soggetti: è precisamente la differenza dei sistemi di rappresentazione — la diversità delle "categorie" cognitivamente rilevanti tra agenti diversi — che rende il malinteso necessario e la comunicazione produttiva. L'universalità kantiana produrrebbe, nel quadro formale, comunicazione perfetta e produttività nulla. La diversità strutturale degli agenti cognitivi produce malinteso inevitabile e produttività positiva.

V.6 Il malinteso come dono: verso una fenomenologia dell'incomprensione produttiva

Si conclude la monografia con una considerazione fenomenologica che il quadro formale giustifica ma non esaurisce. Il malinteso — nella sua versione creativa, non in quella degenerativa o istituzionalizzata come strumento di potere — ha la struttura di un dono asimmetrico: il destinatario riceve dal mittente qualcosa che il mittente non intendeva dare, e questa ricezione non depauperava il mittente ma arricchisce lo spazio condiviso.

Il dono del malinteso è la struttura aggiuntiva DS: la complessità che l'interpretazione del destinatario porta e che l'intenzione originale non aveva. Questa struttura non è arbitraria — è generata dalla specifica configurazione del sistema di rappresentazione del destinatario attivato dal segnale del mittente — e non è senza valore — può essere la sorgente di idee che nessuno dei due possedeva e che entrambi possono usare.

La fenomenologia di questo processo è la fenomenologia della conversazione filosofica autentica, dell'insegnamento creativo, della collaborazione scientifica tra esperti di domini diversi: la sensazione di capire qualcosa che non si aveva capito

prima, non perché il mittente lo abbia spiegato bene, ma perché il proprio sistema di rappresentazione ha generato, a partire dal segnale ricevuto, una struttura che era latente in quel sistema e che il segnale ha attivato. Il destinatario capisce qualcosa che il mittente non sapeva, usando le parole del mittente come chiave per aprire una porta nel proprio sistema cognitivo che il mittente non sapeva che esistesse.

Questo è il limite della comunicazione perfetta come ideale: una comunicazione che trasmettesse esattamente l'intenzione del mittente senza trasformazione non potrebbe mai produrre nel destinatario più di ciò che il mittente già aveva. La comunicazione sarebbe replicazione, non generazione. Il malinteso creativo è la condizione perché la comunicazione sia generativa — perché lo scambio di segnali tra menti diverse produca struttura che non era in nessuna di esse.

V.7 Il limite della tesi: quando il malinteso distrugge

L'onestà intellettuale richiede di non concludere su una nota esclusivamente celebrativa. Il malinteso creativo è produttivo sotto condizioni specifiche che la monografia ha analizzato: sistemi di rappresentazione diversi ma non incompatibili, capacità di produrre $DS > 0$ nel ricevente, contesti in cui la struttura aggiunta è coerente con la struttura profonda del dominio. Fuori da queste condizioni, il malinteso non è dono ma danno.

Il malinteso degenerativo — la distorsione semantica che produce struttura falsa — è la forma più comune del malinteso quotidiano. Il malinteso istituzionalizzato come strumento di potere non arricchisce lo spazio cognitivo collettivo: lo impoverisce selettivamente, erigendo barriere sistematiche all'accesso cognitivo per certi agenti. Il malinteso del modello artificiale — la regressione verso la media del corpus — non amplifica ma smorza: riduce la diversità cognitiva anziché sfruttarla.

La condizione trascendentale del malinteso non è dunque una garanzia di produttività ma una struttura di possibilità: rende possibile la comunicazione produttiva senza renderla necessaria. Realizzare la produttività possibile del malinteso richiede condizioni aggiuntive — la volontà di abitare l'incomprensione senza risolverla prematuramente, la capacità di riconoscere il valore della struttura aggiunta dal destinatario senza ridurla all'intenzione del mittente, l'istituzione di spazi comunicativi in cui il malinteso creativo sia protetto dalla cristallizzazione autoritaria. Queste condizioni non sono formali ma politiche, etiche, istituzionali. Sono il dominio in cui la teoria del malinteso come struttura cognitiva lascia il posto alla pratica del malinteso come responsabilità collettiva.

Capitolo VI — La riparazione conversazionale come protocollo di error-correction

VI.1 Il rilevamento del malinteso come problema di segnalazione

Il quadro formale dei capitoli precedenti ha trattato la trasduzione $\text{Dec}_\beta \circ \text{Enc}_\alpha$ come un processo che produce malinteso senza che nessuno dei due agenti lo percepisca necessariamente. Questo è realistico per il malinteso cristallizzato e per il malinteso artificiale, ma è incompleto per la comunicazione dialogica in tempo reale: in una conversazione, i partecipanti rilevano continuamente segnali di malinteso — esitazioni, richieste di chiarimento, risposte incongruenti — e li usano per correggere la trasduzione in corso. La *riparazione conversazionale* — il meccanismo attraverso cui i partecipanti identificano e correggono i guasti comunicativi — è la controparte dinamica del modello statico della trasduzione.

Il problema formale della riparazione è il seguente. Il destinatario β ha prodotto un'interpretazione $i_\beta = \text{Dec}_\beta(s)$. Il mittente α non ha accesso diretto a i_β : può solo osservare le *azioni* che β compie sulla base di i_β — le sue risposte verbali e non verbali. La distanza di malinteso $\text{MU}(\alpha \rightarrow \beta)$ non è direttamente osservabile da α : deve essere *inferita* dalle azioni di β attraverso un processo di inverse inference — un ragionamento a ritroso dalle azioni osservate all'interpretazione che le ha generate.

Formalmente, α osserva l'azione a_β e deve inferire i_β attraverso il modello:

$$p(i_\beta \mid a_\beta) \propto p(a_\beta \mid i_\beta) \cdot p(i_\beta \mid s)$$

il prodotto del likelihood dell'azione data l'interpretazione e della probabilità a priori dell'interpretazione dato il segnale. Quando $p(i_\beta \mid a_\beta)$ è concentrata su interpretazioni lontane da i_α — quando l'azione di β rivela un'interpretazione distante dall'intenzione — α ha evidenza di malinteso e può attivare la riparazione.

VI.2 La struttura della sequenza di riparazione

Emanuel Schegloff, Harvey Sacks e Gail Jefferson hanno identificato nella conversazione ordinaria una struttura ricorrente della riparazione: la sequenza a tre parti in cui un partecipante inizia la riparazione, l'altro la esegue, e il primo la ratifica o la rifiuta. La sequenza ha varianti — riparazione auto-iniziata dall'emittente, riparazione etero-iniziata dal ricevente, riparazione auto-eseguita o etero-eseguita — e ciascuna variante ha un diverso posizionamento sequenziale preferito nella conversazione.

Nel quadro formale, la sequenza di riparazione è un protocollo di error-correction che opera sullo spazio delle trasduzioni. Si distinguono quattro operazioni elementari.

L'*iniziazione della riparazione* è il segnale con cui un partecipante indica che la trasduzione ha prodotto $\text{MU} > \epsilon$ — che la distanza di malinteso ha superato una soglia tollerabile. Il segnale di iniziazione ha la struttura di un messaggio s_{rep} che trasmette la quantità $\text{KC}(i_\beta \mid i_\alpha)$ stimata da β — il "non ho capito" equivale a $\text{KC}(i_\beta \mid i_\alpha) > 0$ senza specificare la direzione del malinteso; il "non capisco come X possa essere Y" equivale a $\text{KC}(i_\beta \mid i_\alpha) > 0$ con specificazione parziale della discrepanza.

La *specificazione del malinteso* è la comunicazione più dettagliata della discrepanza: il destinatario trasmette informazione sulla struttura di i_β in modo da permettere ad α di stimare la direzione del malinteso. Formalmente, la specificazione riduce l'incertezza di α su i_β :

$$H(i_\beta \mid a_\beta, s_{\text{spec}}) < H(i_\beta \mid a_\beta)$$

dove s_{spec} è il segnale di specificazione e H è l'entropia condizionale.

L'*esecuzione della riparazione* è la modifica di s da parte del mittente — la produzione di un nuovo segnale s' tale che $\text{Dec}_\beta(s')$ sia più vicina all'intenzione i_α di quanto non fosse $\text{Dec}_\beta(s)$. La riparazione è una ricodifica: α sceglie $s' = \text{Enc}'_\alpha(i_\alpha)$ ottimizzato rispetto alla nuova stima di Dec_β prodotta dalla specificazione.

La *ratifica* è la conferma da parte del destinatario che la nuova trasduzione ha ridotto MU sotto la soglia: la risposta che segnala che $i_\beta' = \text{Dec}_\beta(s')$ è sufficientemente vicina a i_α .

VI.3 La dinamica dell'error-correction: convergenza e non convergenza

La sequenza di riparazione è un protocollo iterativo che tenta di convergere verso $\text{MU} = 0$ attraverso successivi affinamenti del segnale. La domanda formale è: sotto quali condizioni converge, e quando non converge?

Si modella la sequenza come un algoritmo di ottimizzazione su s . Ad ogni iterazione t , il mittente produce un segnale $s^\wedge\{t\}$ e osserva il segnale di riparazione $r^\wedge\{t\}$ del destinatario. Il segnale di riparazione fornisce informazione sul gradiente di MU rispetto a s : indica in quale direzione modificare s per ridurre la distanza di malinteso. Il mittente aggiorna s nella direzione indicata:

$$s^\wedge\{t+1\} = s^\wedge\{t\} - \eta_{\text{rep}} \cdot \nabla_s \text{MU}(\alpha \rightarrow \beta; s^\wedge\{t\})$$

dove η_{rep} è il tasso di aggiornamento della riparazione e $\nabla_s \text{MU}$ è il gradiente della distanza di malinteso rispetto al segnale.

La convergenza di questo algoritmo è garantita sotto due condizioni. La prima è che $\nabla_s \text{MU}$ sia calcolabile — che il segnale di riparazione $r^\wedge\{t\}$ fornisca informazione sufficiente per stimare la direzione del gradiente. La seconda è che $\text{MU}(\alpha \rightarrow \beta; s)$ sia una funzione convessa di s — che esista un segnale globalmente ottimale s^* verso cui il gradiente punta.

La prima condizione fallisce quando il destinatario non può specificare la propria interpretazione errata in modo sufficientemente informativo. Questo accade nei malintesi profondi — malintesi in cui la discrepanza è a livello delle strutture portanti di M_β , non dei dettagli periferici — perché la discrepanza a quel livello non è

accessibile all'introspezione del destinatario. Il destinatario sa che qualcosa non torna ma non riesce a indicare dove: la sua specificazione del malinteso è insufficientemente informativa per guidare la riparazione.

La seconda condizione fallisce quando il malinteso è strutturalmente irriducibile — quando la distanza tra M_α e M_β è tale che nessun segnale s può produrre $\text{Dec}_\beta(s) = i_\alpha$. In questo caso $\text{MU}(\alpha \rightarrow \beta; s)$ ha un minimo globale positivo: anche la riparazione ottimale lascia un residuo di malinteso non eliminabile. Questo è il caso delle comunicazioni tra esperti di domini radicalmente diversi, delle controversie filosofiche su questioni fondamentali, delle negoziazioni tra agenti con sistemi di valori incompatibili.

VI.4 Il costo della riparazione e la soglia di tolleranza

La riparazione ha un costo: richiede ulteriori scambi comunicativi, introduce latenza nel flusso della conversazione, può produrre imbarazzo o conflitto. Il mittente che decide se avviare la riparazione deve bilanciare il costo del malinteso irrisolto con il costo della riparazione stessa.

Si definisce la *soglia di riparazione* ϵ_{rep} come il valore minimo di MU che giustifica l'avvio della riparazione. Il mittente avvia la riparazione se e solo se:

$$\text{MU}(\alpha \rightarrow \beta) > \epsilon_{\text{rep}} \quad \text{e} \quad E[\Delta \text{MU} \mid \text{riparazione}] > \text{Costo}_{\text{rep}}$$

dove $E[\Delta \text{MU} \mid \text{riparazione}]$ è la riduzione attesa del malinteso per effetto della riparazione e $\text{Costo}_{\text{rep}}$ è il costo della sequenza di riparazione. La soglia ϵ_{rep} è modulata da fattori contestuali — la posta in gioco comunicativa, la relazione tra i partecipanti, le norme conversazionali del contesto — che determinano sia il costo del malinteso irrisolto che il costo della riparazione.

Questa analisi predice sistematicamente dove il malinteso si cristallizza anziché essere corretto: nei contesti in cui il $\text{Costo}_{\text{rep}}$ è artificialmente elevato — relazioni di potere asimmetrica in cui il subordinato non può contraddire il superiore, contesti rituali in cui la fluidità conversazionale ha valore superiore alla precisione comunicativa, situazioni in cui ammettere il malinteso è percepito come perdita di faccia. In questi contesti, i malintesi superano la soglia ϵ_{rep} ma la sequenza di riparazione non viene attivata: il malinteso persiste, si accumula, e nel tempo si cristallizza nella forma descritta nel Capitolo III.

VI.5 La meta-comunicazione come riparazione di secondo ordine

Esiste un livello di riparazione superiore a quello delle sequenze conversazionali elementari: la *meta-comunicazione*, la comunicazione sulla comunicazione. Quando la riparazione di primo ordine — la modifica del segnale s — non converge, i partecipanti possono tentare di riparare il processo di trasduzione stesso: invece di modificare il segnale all'interno del sistema di decodifica di β , tentano di modificare il

sistema di decodifica di β — di installare in β nuove primitive che riducano $MU_{sys}(\alpha, \beta)$.

Formalmente, la meta-comunicazione è un'operazione sul sistema di rappresentazione del destinatario:

$$M_{\beta'} = M_{\beta} + \Delta M$$

dove ΔM è la modifica del sistema di rappresentazione prodotta dalla meta-comunicazione. La riparazione di secondo ordine riduce $MU_{sys}(\alpha, \beta') < MU_{sys}(\alpha, \beta)$ modificando il sistema di decodifica piuttosto che il segnale.

Le forme di meta-comunicazione includono la definizione dei termini — che installa in M_{β} le primitive necessarie per decodificare i_{α} correttamente — la spiegazione del contesto presupposto — che fornisce a M_{β} la struttura di background che il segnale dava per scontata — e la costruzione di analogie — che collega le primitive assenti in M_{β} a primitive presenti, riducendo il gap di simulazione del Capitolo VI della prima monografia.

La meta-comunicazione è più costosa della riparazione di primo ordine — modifica il sistema di rappresentazione anziché il segnale — ma è l'unico meccanismo disponibile per ridurre MU_{sys} quando la distanza strutturale tra i sistemi di rappresentazione è la sorgente del malinteso. La buona insegnamento è meta-comunicazione sistematica: non trasmissione di contenuto ma costruzione delle primitive necessarie per ricevere il contenuto.

VI.6 Il fallimento della riparazione come dato diagnostico

Il fallimento sistematico della riparazione — la persistenza del malinteso nonostante tentativi ripetuti di correzione — è un dato diagnostico preciso sulla struttura del malinteso. Il quadro formale predice due cause distinte di fallimento persistente.

Il fallimento per *inaccessibilità del gradiente* — il destinatario non riesce a specificare la propria interpretazione errata — indica un malinteso a livello delle strutture portanti di M_{β} : la discrepanza è talmente profonda che non è accessibile all'introspezione. La sequenza di riparazione produce segnali di specificazione sempre più vaghi — "non capisco", "non ci arrivo", "non riesco a vederlo" — senza mai identificare la struttura dell'errore. Il rimedio è la meta-comunicazione profonda: costruire le primitive mancanti piuttosto che correggere il segnale.

Il fallimento per *incompatibilità strutturale* — la riparazione riduce MU ma non sotto la soglia accettabile — indica un malinteso irriducibile: la distanza tra M_{α} e M_{β} è tale che $MU_{sys} > 0$ è inevitabile per quel contenuto. La sequenza di riparazione converge a un minimo positivo — il malinteso residuo minimo compatibile con la distanza tra i sistemi. Il rimedio non è più correggere il segnale né modificare i sistemi individuali: è costruire un terzo sistema di rappresentazione condiviso — un pidgin cognitivo, nel

senso del Capitolo VII della monografia sulla comunicazione a due agenti — che riduca il gap senza richiedere che nessuno dei due agenti abbandoni il proprio sistema.

Capitolo VII — Il malinteso intrapersonale: autoinganno, razionalizzazione e deriva del sé

VII.1 Il caso limite: mittente e destinatario coincidono

Il quadro formale della trasduzione presuppone due agenti distinti — un mittente α e un destinatario β con sistemi di rappresentazione M_α e M_β . Ma il quadro si estende naturalmente a un caso limite in cui mittente e destinatario coincidono: l'agente che codifica un'intenzione, un'emozione, un ricordo in un segnale interno — linguaggio interiore, immagine mentale, schema motorio — e poi lo decodifica in uno stato cognitivo che può differire dall'intenzione originale.

Questo caso non è una curiosità marginale. È la struttura dell'autoinganno, della razionalizzazione post hoc, del lapsus, della dissonanza cognitiva, dell'insight che rivela che ciò che si credeva di volere non era ciò che si voleva davvero. Sono tutti fenomeni in cui l'agente produce una rappresentazione di uno stato interno e poi quella rappresentazione non corrisponde allo stato che l'ha generata — in cui $\text{Dec}_{\{M_{\{t_1\}}\}}(\text{Enc}_{\{M_{\{t_0\}}\}}(i)) \neq i$ per la stessa mente a tempi diversi.

Il meccanismo formale è la *deriva del sistema di rappresentazione nel tempo*. La monografia sulla memoria ha dimostrato che $m_e^{\{T\}} \neq m_e^{\{0\}}$ — che il modello mnestico deriva attraverso il riconsolidamento. Lo stesso meccanismo si applica al sistema di rappresentazione globale M_t : il programma del sé $p_{\{s\}}^{\{t\}}$ deriva nel tempo attraverso gli aggiornamenti bayesiani del Capitolo II di quella monografia. Quando $M_{\{t_0\}}$ e $M_{\{t_1\}}$ differiscono — quando il sistema di rappresentazione dell'agente è cambiato tra il momento della codifica e il momento della decodifica — la trasduzione intrapersonale $\text{Dec}_{\{M_{\{t_1\}}\}} \circ \text{Enc}_{\{M_{\{t_0\}}\}}$ non è l'identità.

VII.2 Formalizzazione del malinteso intrapersonale

Si definisce il *malinteso intrapersonale* come la distanza di malinteso nella trasduzione intrapersonale:

$$\text{MU}_{\text{intra}}(t_0, t_1; i) = \text{KC}(i \mid \text{Dec}_{\{M_{\{t_1\}}\}}(\text{Enc}_{\{M_{\{t_0\}}\}}(i)))$$

la complessità condizionale dell'intenzione originale i rispetto all'interpretazione che l'agente produce di se stesso al tempo t_1 . Il malinteso intrapersonale è zero quando $M_{\{t_0\}} = M_{\{t_1\}}$ — quando il sistema di rappresentazione non è derivato — e positivo quando la deriva ha portato i due sistemi a distanza strutturale.

La distorsione semantica intrapersonale è:

$$DS_{intra}(t_0, t_1; i) = KC(Dec_{\{M_{\{t_1\}}\}}(Enc_{\{M_{\{t_0\}}\}}(i)) \mid i)$$

la struttura che l'agente al tempo t_1 aggiunge all'intenzione originale. $DS_{intra} > 0$ indica che l'agente interpreta i propri stati passati con una ricchezza che al momento dell'intenzione non possedeva: legge struttura nei propri ricordi che la struttura originale non aveva. Questo è il meccanismo della *costruzione retrospettiva del significato* — l'esperienza che il senso di un evento diventa chiaro solo molto tempo dopo che l'evento è accaduto, quando il sistema di rappresentazione si è sviluppato abbastanza da vedere ciò che allora era opaco.

VII.3 La razionalizzazione come malinteso intrapersonale sistematico

La razionalizzazione — il processo attraverso cui l'agente produce spiegazioni post hoc delle proprie azioni che non corrispondono alle cause effettive — è un caso di malinteso intrapersonale con struttura specifica. L'azione è prodotta da processi cognitivi che operano nel sistema $M_{\{t_0\}}$ — un sistema che include stati subconsci, automatismi, valutazioni affettive non verbalizzate. La spiegazione è prodotta da processi nel sistema $M_{\{t_1\}}$ — un sistema che include le primitive verbali e narrative dell'agente al momento della riflessione.

La trasduzione $Dec_{\{M_{\{t_1\}}\}}(Enc_{\{M_{\{t_0\}}\}}(i))$ mappa la causa effettiva dell'azione — che opera in $M_{\{t_0\}}$ — nell'interpretazione prodotta dal sistema narrativo $M_{\{t_1\}}$. Quando $M_{\{t_0\}}$ e $M_{\{t_1\}}$ differiscono — quando il sistema subcoscio che ha generato l'azione e il sistema narrativo che la spiega sono strutturalmente distanti — la trasduzione produce $MU_{intra} > 0$: la spiegazione non corrisponde alla causa.

Questa è la struttura formale della razionalizzazione nel senso di Nisbett e Wilson (1977): l'agente non ha accesso diretto ai propri processi cognitivi subconsci e produce spiegazioni che sono plausibili nel proprio sistema narrativo $M_{\{t_1\}}$ ma non corrispondono ai processi in $M_{\{t_0\}}$ che hanno effettivamente generato il comportamento. La razionalizzazione non è menzogna deliberata: è inferenza bayesiana sotto un sistema di rappresentazione sbagliato.

Formalmente, la spiegazione razionalizzata è:

$$i_{\{exp\}} = \operatorname{argmax}_{\{i'\}} p_{\{M_{\{t_1\}}\}}(i' \mid \text{azione, contesto})$$

l'interpretazione più a priori plausibile nel sistema narrativo dell'agente data l'azione e il contesto. Questa interpretazione massimizza la coerenza narrativa nel sistema $M_{\{t_1\}}$ — è la spiegazione più "sensata" nel sistema di rappresentazione corrente — ma può essere lontana dalla causa effettiva i che ha operato in $M_{\{t_0\}}$.

VII.4 L'autoinganno come malinteso intrapersonale stabilizzato

L'autoinganno è una forma di malinteso intrapersonale più profonda della razionalizzazione: non soltanto la produzione post hoc di spiegazioni false ma la costruzione attiva di interpretazioni di sé che contraddicono l'evidenza disponibile. Il soggetto che si autoinganna non solo non sa perché fa ciò che fa, ma attivamente resiste all'evidenza che contraddirebbe la propria interpretazione di sé.

Nel quadro formale, l'autoinganno è un malinteso intrapersonale con meccanismo di stabilizzazione analogo a quello del ricordo traumatico nel Capitolo II della monografia sulla memoria. La rappresentazione di sé i_false ha alta valenza positiva per il programma del sé $p_sé$ — rappresenta l'agente nel modo che il sé narrativo preferisce — e quindi il tasso di aggiornamento η verso interpretazioni che contraddicono i_false è ridotto:

$$\eta(i_false, t) = \eta_0 \cdot \exp(-\lambda \cdot v(i_false))$$

dove $v(i_false) > 0$ è la valenza positiva della rappresentazione falsa. L'aggiornamento verso la verità è lento, e può essere superato dagli aggiornamenti che confermano la rappresentazione falsa attraverso il riconsolidamento in contesti congruenti.

La previsione formale è che l'autoinganno più stabile riguardi le rappresentazioni di sé con più alta valenza — quelle che toccano i tratti del sé più centrali all'identità narrativa. L'agente si autoinganna più facilmente sulle proprie motivazioni morali, sulla propria intelligenza, sul proprio coraggio — tratti centrali al sé narrativo — che su fatti periferici. La letteratura sulla self-serving bias è qualitativamente consistente con questa previsione.

VII.5 Il lapsus come errore della trasduzione intrapersonale

Il lapsus freudiano — il lapsus linguae, il lapsus calami, l'atto mancato — ha nel quadro formale una struttura precisa come errore della trasduzione intrapersonale. Il lapsus è la situazione in cui la codifica $Enc_{\{M_{\{t_0\}}\}}$ produce un segnale s che è distante dal segnale che il sistema narrativo $M_{\{t_1\}}$ avrebbe prodotto per la stessa intenzione — in cui un'intenzione subconscia si infila nella codifica del segnale.

Si modella il processo di codifica come la sovrapposizione di due sistemi: il sistema narrativo conscio M_{cons} che genera il segnale inteso s_{int} , e il sistema subconscio M_{sub} che genera il segnale subinteso s_{sub} . Il segnale effettivamente prodotto è:

$$s = (1-\alpha) \cdot s_{int} + \alpha \cdot s_{sub} + \epsilon$$

dove $\alpha \in [0,1]$ misura il peso relativo del sistema subconscio nella codifica e ϵ è rumore. Per $\alpha \approx 0$, il segnale è quello inteso: il sistema conscio domina. Per $\alpha > 0$, il segnale è influenzato dal sistema subconscio: il lapsus emerge come la componente $\alpha \cdot (s_{sub} - s_{int})$ — la differenza tra il segnale subconscio e quello conscio, pesata per il peso del sistema subconscio.

La struttura del lapsus riflette la struttura della discrepanza tra M_{cons} e M_{sub} . Lapsus che sostituiscono parole con parole semanticamente correlate riflettono discrepanze nei nodi semantici — il sistema subcoscio ha attivato un concetto associato. Lapsus che invertono l'ordine delle parole riflettono discrepanze nella struttura sequenziale. Lapsus che producono parole con valenza emotiva opposta riflettono discrepanze nella dimensione valutativa del sistema di rappresentazione.

VII.6 L'insight come riparazione del malinteso intrapersonale

Il quadro formale permette di reinterpretare l'insight — il fenomeno del Capitolo II della monografia sul pensiero come compressione — come riparazione del malinteso intrapersonale. L'insight è il momento in cui la trasduzione intrapersonale $\text{Dec}_{\{M_{\{t_1\}}\}} \circ \text{Enc}_{\{M_{\{t_0\}}\}}$ passa da produrre MU_{intra} elevata a $\text{MU}_{\text{intra}} \approx 0$: il sistema di rappresentazione attuale $M_{\{t_1\}}$ diventa abbastanza sviluppato da decodificare fedelmente ciò che il sistema originale $M_{\{t_0\}}$ aveva codificato.

Questo reinterpreta l'insight non come la comparsa di qualcosa di nuovo — una comprensione prima assente — ma come la rimozione di un malinteso intrapersonale: la struttura era già codificata in $M_{\{t_0\}}$, ma il sistema di rappresentazione al momento della riflessione non era sufficientemente sviluppato da decodificarla correttamente. Quando $M_{\{t_1\}}$ ha acquisito le primitive necessarie — attraverso apprendimento, esperienza, maturazione — la decodifica retroattiva produce la sensazione improvvisa di aver "capito finalmente" qualcosa che era sempre stato lì.

Questa reinterpretazione ha una previsione verificabile: l'insight dovrebbe essere più frequente in agenti che hanno recentemente acquisito nuove primitive nel proprio sistema di rappresentazione — chi ha appena imparato un nuovo dominio è più soggetto a insight retrospettivi su esperienze passate che ora diventano comprensibili. La letteratura sulla "retrospective sense-making" in psicologia è consistente con questa previsione, sebbene raramente la formalizzi in questi termini.

VII.7 La connessione con la memoria come finzione: il sé che si malintende

Il malinteso intrapersonale connette questa monografia con quella sulla memoria come finzione in modo che non era stato esplicitato in nessuna delle due. La deriva del programma del sé $p_{\{\text{sé}\}^{\{t\}}}$ — il meccanismo centrale del Capitolo II della monografia sulla memoria — è precisamente il meccanismo che genera la distanza tra $M_{\{t_0\}}$ e $M_{\{t_1\}}$: la fonte del malinteso intrapersonale.

Un agente il cui $p_{\{\text{sé}\}^{\{t\}}}$ ha derivato significativamente da $p_{\{\text{sé}\}^{\{0\}}}$ malintende sistematicamente i propri stati passati: li decodifica nel sistema $M_{\{t_1\}}$ producendo interpretazioni che sono lontane da ciò che l'agente a t_0 intendeva. Il ricordo distorto del Capitolo I della monografia sulla memoria è, in questa lettura, un malinteso intrapersonale: l'agente al tempo t_1 malintende l'agente al tempo t_0 , esattamente come due agenti con sistemi di rappresentazione diversi si malintendono nel caso interpersonale.

La differenza strutturale tra il malinteso intrapersonale e quello interpersonale è che nel caso intrapersonale non esiste un mittente e un destinatario separati che possano avviare una sequenza di riparazione. L'agente non può comunicare con se stesso al tempo t_0 per correggere la trasduzione: il passato non è accessibile. L'unica riparazione disponibile è la riparazione di secondo ordine — la meta-comunicazione che modifica $M_{\{t_1\}}$ nella direzione di $M_{\{t_0\}}$: lo sforzo di recuperare i sistemi di rappresentazione del passato attraverso la documentazione, il ricordo deliberato, la ricostruzione storica del contesto in cui i propri stati passati furono generati.

Questa connessione chiude il cerchio tra le due monografie: la memoria come finzione è il sedimento del malinteso intrapersonale accumulato nel tempo — il deposito di tutte le trasduzione $Dec_{\{M_{\{t\}}\}} \circ Enc_{\{M_{\{t_0\}}\}}$ che l'agente ha eseguito su se stesso nel corso della propria storia. E il sé narrativo — il programma del sé $p_{\{sé\}}$ — è al tempo stesso il generatore delle trasduzione intrapersonali e il loro prodotto: è la struttura che si è formata attraverso il sedimento di tutti i malintesi con se stesso, e che genera i malintesi futuri.

Appendice A — Dimostrazione del teorema dell'impossibilità della comunicazione perfetta

A.1 Enunciato preciso

Siano M_α e M_β due sistemi di rappresentazione con insiemi di primitive P_α e P_β . La trasduzione $Dec_\beta \circ Enc_\alpha$ soddisfa:

$$\begin{aligned} MU(\alpha \rightarrow \beta; i) &= KC(i \mid Dec_\beta(Enc_\alpha(i))) = 0 \\ DS(\alpha \rightarrow \beta; i) &= KC(Dec_\beta(Enc_\alpha(i)) \mid i) = 0 \end{aligned}$$

per ogni intenzione i non triviale se e solo se $M_\alpha = M_\beta$.

A.2 Dimostrazione

Direzione necessaria ($M_\alpha = M_\beta \Rightarrow MU = DS = 0$).

Se $M_\alpha = M_\beta$, allora Enc_α e Dec_β sono costruite sullo stesso insieme di primitive $P = P_\alpha = P_\beta$. La funzione Dec_β è per costruzione l'inversa di Enc_α sullo stesso spazio: $Dec_\beta(Enc_\alpha(i)) = i$ per ogni i nel dominio della codifica. Dunque $KC(i \mid Dec_\beta(Enc_\alpha(i))) = KC(i \mid i) = 0$ e $KC(Dec_\beta(Enc_\alpha(i)) \mid i) = KC(i \mid i) = 0$.

Direzione sufficiente ($MU = DS = 0 \Rightarrow M_\alpha = M_\beta$).

$MU(\alpha \rightarrow \beta; i) = 0$ implica $KC(i \mid Dec_\beta(Enc_\alpha(i))) = 0$, ossia $i_\beta = Dec_\beta(Enc_\alpha(i))$ determina completamente i : esiste un programma di lunghezza $O(1)$ che ricostruisce i da i_β . $DS(\alpha \rightarrow \beta; i) = 0$ implica $KC(i_\beta \mid i) = 0$, ossia i determina completamente i_β .

Le due condizioni insieme implicano che i e i_β si determinano a vicenda in $O(1)$ — sono computazionalmente equivalenti a meno di una costante. Per la definizione di trasduzione, $Dec_\beta(Enc_\alpha(i)) = i$ a meno di $O(1)$ bit per ogni i . Questo richiede che Enc_α sia invertibile rispetto a Dec_β con overhead costante, cioè che $Dec_\beta \circ Enc_\alpha = id + O(1)$. Per il teorema di invarianza (Appendice A della prima monografia), questo implica che M_α e M_β simulano l'una l'altra a costo costante — condizione che, per macchine di rappresentazione parziali nel senso del Capitolo VII della prima monografia, equivale a $P_\alpha = P_\beta$, cioè $M_\alpha = M_\beta$. ■

A.3 Corollario sulla non-simmetria

Il teorema implica che $MU(\alpha \rightarrow \beta)$ e $MU(\beta \rightarrow \alpha)$ sono in generale diversi. Si verifica:

Se $P_\beta \supset P_\alpha$ — il sistema di β ha tutte le primitive di α più altre aggiuntive — allora Enc_α produce segnali che Dec_β può decodificare fedelmente usando le primitive condivise $P_\alpha \subseteq P_\beta$, ma i segnali $Enc_\beta(i)$ usano primitive di $P_\beta \setminus P_\alpha$ che Dec_α non può decodificare correttamente. Dunque $MU(\alpha \rightarrow \beta) \approx 0$ mentre $MU(\beta \rightarrow \alpha) > 0$: la comunicazione dall'esperto al novizio è più problematica della comunicazione dal novizio all'esperto, in accordo con l'osservazione pedagogica comune.

Appendice B — La funzione di tasso-distorsione comunicativa

B.1 Definizione del problema rate-distortion per la comunicazione

Il problema di ottimizzazione della codifica comunicativa è: dato un sistema di rappresentazione M_α del mittente e M_β del destinatario, trovare la funzione di codifica Enc_α^* che minimizza la lunghezza del segnale sotto il vincolo che la distorsione di malinteso non superi D_{target} :

$$Enc_\alpha^* = \operatorname{argmin}_{\{Enc_\alpha\}} \mathbb{E}[|Enc_\alpha(i)|] \\ \text{soggetto a } \mathbb{E}[MU(\alpha \rightarrow \beta; i)] \leq D_{target}$$

La funzione di tasso-distorsione comunicativa $R_{comm}(D)$ è la lunghezza minima del segnale attesa compatibile con distorsione media D :

$$R_{comm}(D) = \min_{\{P(s|i) : \mathbb{E}[MU] \leq D\}} I(I; S)$$

dove I è la variabile aleatoria sull'intenzione, S è la variabile aleatoria sul segnale, e $I(I; S)$ è la loro informazione mutua.

B.2 Il lower bound di Shannon per la comunicazione cognitiva

Per il teorema di Shannon sulla compressione con perdita applicato alla comunicazione:

$$|Enc_{\alpha}(i)| \geq R_{comm}(D_{target})$$

con probabilità che tende a 1 per grandi blocchi di intenzioni. Qualsiasi schema di codifica che garantisca distorsione di malinteso media D_{target} deve usare segnali di lunghezza almeno $R_{comm}(D_{target})$ bit in media.

Poiché $MU(\alpha \rightarrow \beta; i)$ include la componente sistematica $MU_{sys}(\alpha, \beta) \geq 0$, si ha:

$$D_{target} \geq MU_{sys}(\alpha, \beta) \Rightarrow R_{comm}(D_{target}) \geq 0$$

$$D_{target} < MU_{sys}(\alpha, \beta) \Rightarrow R_{comm}(D_{target}) = \infty$$

La seconda condizione formalizza il fatto che se la distorsione target è inferiore al malinteso sistematico irriducibile, nessuno schema di codifica può raggiungerla: la comunicazione con distorsione sub-sistematica è impossibile indipendentemente dalla lunghezza del segnale.

B.3 La massima di quantità come vincolo rate-distortion

La massima di quantità di Grice richiede che il segnale non sia né troppo lungo né troppo breve. Nel quadro rate-distortion:

Il vincolo di non-brevità è il lower bound $R_{comm}(D_{target})$: il segnale deve essere almeno lungo quanto richiesto dalla funzione di tasso-distorsione per garantire la distorsione target.

Il vincolo di non-ridondanza richiede che il segnale non superi significativamente $R_{comm}(D_{target})$: la ridondanza aggiuntiva non riduce la distorsione di malinteso ma può aumentare il rumore fisico sul canale, la probabilità che il destinatario perda elementi del segnale, e il costo cognitivo dell'elaborazione.

La violazione per eccesso — segnali più lunghi del necessario — produce un effetto paradossale: aumenta la probabilità che il destinatario presti attenzione alle parti ridondanti anziché alle parti informative, producendo una distribuzione di MU peggiore di quella che un segnale più conciso avrebbe prodotto.

Appendice C — Il modello formale della fertilità delle idee

C.1 La catena di Markov delle trasduzione

Si modella la catena di trasduzione $\alpha_1 \rightarrow \alpha_2 \rightarrow \dots \rightarrow \alpha$ come una catena di Markov sui contenuti delle idee. Lo stato al tempo k è l'idea i_k , e la matrice di transizione è determinata dalla trasduzione $\text{Dec}_{\{M_{k+1}\}} \circ \text{Enc}_{\{M_k\}}$:

$$P(i_{k+1} = j \mid i_k = i) = P(\text{Dec}_{\{M_{k+1}\}}(\text{Enc}_{\{M_k\}}(i)) = j)$$

La distribuzione stazionaria π della catena — se esiste — descrive la distribuzione delle idee che la catena produce nel lungo termine, indipendentemente dall'idea iniziale.

C.2 La fertilità come divergenza dalla distribuzione stazionaria

La fertilità di un'idea originale i_0 rispetto alla catena è:

$$\begin{aligned} \text{Fert}_n(i_0) &= \mathbb{E}[\text{KC}(i)] - \text{KC}(i_0) \\ &= \sum_j \pi_n(j \mid i_0) \cdot \text{KC}(j) - \text{KC}(i_0) \end{aligned}$$

dove $\pi_n(j \mid i_0)$ è la distribuzione dell'idea al passo n data l'idea iniziale i_0 . La fertilità è positiva se la catena di trasduzione produce idee con KC media superiore a quella dell'idea originale — se il malinteso creativo amplifica struttura.

C.3 La condizione di amplificazione

La condizione $\text{KC}(i_{k+1}) > \text{KC}(i_k)$ in media richiede che la trasduzione $\text{Dec}_{\{M_{k+1}\}} \circ \text{Enc}_{\{M_k\}}$ sia in media espansiva — che produca $\text{DS} > \text{MU}$ in attesa. Si può dimostrare che questa condizione è soddisfatta quando:

$$\mathbb{E}_{\{i_k\}}[\text{KC}(\text{Ctx}_{\{M_{k+1}\}}(\text{Enc}_{\{M_k\}}(i_k)))] > \mathbb{E}_{\{i_k\}}[\text{MU}_{\text{sys}}(M_k, M_{k+1})]$$

la complessità media del contesto attivato nel sistema del successore supera il malinteso sistematico medio tra i sistemi successivi. Quando i sistemi di rappresentazione sono abbastanza diversi da attivare contesti ricchi ma non così diversi da produrre incompatibilità strutturale, la catena è amplificativa.

Appendice D — La derivazione della soglia ottimale di riparazione

D.1 Il problema di ottimizzazione

Il mittente decide se avviare la riparazione confrontando il beneficio atteso con il costo. Sia $V(\text{MU})$ il costo del malinteso irrisolto in funzione della distanza MU — una funzione crescente in MU. Sia C_{rep} il costo fisso della sequenza di riparazione. Sia $E[\Delta\text{MU} \mid \text{rep}]$ la riduzione attesa del malinteso per effetto della riparazione.

Il mittente avvia la riparazione se e solo se:

$$V(\text{MU}) - V(\text{MU} - E[\Delta\text{MU} \mid \text{rep}]) > C_{\text{rep}}$$

Supponendo V lineare — $V(\text{MU}) = c_{\text{MU}} \cdot \text{MU}$ per una costante $c_{\text{MU}} > 0$ — la condizione diventa:

$$c_{\text{MU}} \cdot E[\Delta\text{MU} \mid \text{rep}] > C_{\text{rep}}$$

La soglia di riparazione è il valore minimo di MU per cui questa condizione è soddisfatta:

$$\epsilon_{\text{rep}} = C_{\text{rep}} / (c_{\text{MU}} \cdot (\partial E[\Delta\text{MU} \mid \text{rep}] / \partial \text{MU}))$$

D.2 Modulazione contestuale della soglia

La soglia ϵ_{rep} è modulata da tre fattori contestuali.

Il costo C_{rep} varia con il contesto comunicativo. In relazioni gerarchiche, C_{rep} include il rischio di perdita di faccia e il costo della violazione della norma di non-contraddizione del superiore: C_{rep} è alto, e ϵ_{rep} è alta — il subordinato tollera malintesi grandi prima di avviare la riparazione. In relazioni simmetriche tra pari, C_{rep} è basso, e ϵ_{rep} è bassa — i partecipanti avviano la riparazione per malintesi piccoli.

La costante c_{MU} varia con la posta in gioco comunicativa. In contesti in cui l'accuratezza è critica — comunicazione medica, negoziazione legale — c_{MU} è alta e ϵ_{rep} è bassa. In contesti in cui la fluidità conversazionale è più importante dell'accuratezza — small talk, narrazione informale — c_{MU} è bassa e ϵ_{rep} è alta.

La riduzione attesa $E[\Delta\text{MU} \mid \text{rep}]$ varia con la calcolabilità del gradiente. Per malintesi superficiali con gradiente accessibile, $E[\Delta\text{MU} \mid \text{rep}]$ è alta: la riparazione è probabile che produca riduzione significativa. Per malintesi profondi con gradiente inaccessibile,

$E[\Delta MU \mid \text{rep}]$ è bassa: la riparazione è improbabile che converga, e e_{rep} sale — il partecipante tollera il malinteso profondo anziché tentare una riparazione improbabile.

Appendice E — Il modello di composizione dei malintesi artificiali

E.1 La composizione di n trasduzione artificiali

Si analizza la composizione di n sistemi di linguaggio in serie. Sia $M_{\{\text{LLM}\}}$ il sistema di rappresentazione implicito di ciascun modello (per semplicità si assume uguale per tutti i nodi). La composizione produce:

$$i_n = \text{Dec}_{\{\text{LLM}\}}^{\{(n)\}}(\text{Enc}_{\{\text{LLM}\}}^{\{(n-1)\}}(\dots \text{Dec}_{\{\text{LLM}\}}^{\{(1)\}}(\text{Enc}_{\alpha}(i_0)) \dots))$$

E.2 La convergenza verso la media del corpus

Si dimostra che per $n \rightarrow \infty$ la distribuzione di i_n converge alla distribuzione stazionaria della catena di Markov indotta dalla trasduzione artificiale, indipendentemente da i_0 .

La trasduzione $\text{Dec}_{\{\text{LLM}\}} \circ \text{Enc}_{\{\text{LLM}\}}$ ha la struttura di un operatore di proiezione verso la distribuzione media del corpus: contrae le idee verso $M_{\{\text{LLM}\}}$. Per il teorema del punto fisso di Banach, se la trasduzione è una contrazione — se $d(\text{Dec}_{\{\text{LLM}\}}(\text{Enc}_{\{\text{LLM}\}}(i)), \text{Dec}_{\{\text{LLM}\}}(\text{Enc}_{\{\text{LLM}\}}(j))) < \lambda \cdot d(i, j)$ per $\lambda < 1$ — allora la composizione iterata converge al punto fisso $i^* = \text{Dec}_{\{\text{LLM}\}}(\text{Enc}_{\{\text{LLM}\}}(i^*))$, che è il centro di massa della distribuzione del corpus.

Il rate di convergenza è:

$$d(i_n, i^*) \leq \lambda^n \cdot d(i_0, i^*)$$

esponenzialmente rapido nel numero di trasduzione. Per n moderato ($n = 5-10$), la maggior parte dell'informazione specifica di i_0 è già persa se i_0 è lontano da i^* .

E.3 Il mode collapse come conseguenza strutturale

Il mode collapse — la riduzione della diversità degli output in sistemi multi-agente — è una conseguenza diretta della convergenza verso i^* . Idee inizialmente diverse $i_0^{\{(1)\}}$, $i_0^{\{(2)\}}$, ..., $i_0^{\{(m)\}}$ che vengono elaborate dalla stessa catena di trasduzione artificiali convergono tutte verso i^* , producendo output indistinguibili indipendentemente dalla diversità degli input.

La velocità del mode collapse è proporzionale al tasso di contrazione λ — che è determinato dalla concentrazione della distribuzione del corpus. Corpus concentrati (poca diversità di sistemi di rappresentazione) producono λ vicino a 1 e convergenza lenta ma inevitabile. Corpus dispersi producono λ più piccolo e convergenza più rapida, ma il punto fisso i^* è meno definito — il sistema oscilla intorno a una regione dell'idea-spazio anziché convergere a un punto.

Bibliografia ragionata

La bibliografia è organizzata per nuclei tematici corrispondenti ai movimenti principali della monografia. Ogni testo è accompagnato da una nota che specifica il ruolo nell'argomentazione e, dove necessario, la distanza critica che la monografia mantiene.

I. Teoria della comunicazione e semantica formale

Shannon, C. E., & Weaver, W. (1949). *The Mathematical Theory of Communication*. University of Illinois Press.

Il testo fondativo che questa monografia sistematicamente supera. Shannon e Weaver modellano la comunicazione come trasmissione di simboli su un canale con rumore, presupponendo un codice condiviso tra mittente e destinatario. Il capitolo di Weaver — più accessibile e filosoficamente ambizioso di quello di Shannon — anticipa il problema della comunicazione tra sistemi con significati diversi ma lo subordina al problema tecnico del canale. La monografia prende le mosse esattamente dove Shannon e Weaver si fermano: quando il "codice condiviso" non è dato ma è esso stesso il prodotto di una storia di trasduzione.

Grice, H. P. (1975). Logic and conversation. In P. Cole & J. Morgan (Eds.), *Syntax and Semantics, Vol. 3: Speech Acts* (pp. 41–58). Academic Press.

L'articolo in cui Grice introduce il principio di cooperazione e le massime conversazionali. La formalizzazione del Capitolo I.6 reinterpreta le massime come vincoli sul codificatore ottimale per la funzione di tasso-distorsione comunicativa. La monografia si distanzia da Grice su un punto preciso: Grice tratta le implicature come deviazioni dai casi normali di comunicazione letterale, mentre la monografia tratta le implicature come uso intenzionale della struttura del malinteso sistematico — non deviazione dalla norma ma sfruttamento strutturale della trasduzione.

Sperber, D., & Wilson, D. (1986/1995). *Relevance: Communication and Cognition* (2^a ed.). Blackwell.

La teoria della pertinenza è il progetto più elaborato di formalizzare la comunicazione come processo inferenziale basato sulla minimizzazione del costo cognitivo. Sperber e Wilson argomenta che il destinatario interpreta il messaggio come l'intenzione che massimizza gli effetti cognitivi minimizzando il costo di elaborazione. Questa è una versione specifica della decodifica ottimale $\text{Dec}_\beta(s) = \text{argmax}_{\{i\}} [\text{effetti}(i) - \text{costo}(i|s)]$ — la decodifica che massimizza il rapporto beneficio-costi nel sistema di rappresentazione del destinatario. La monografia incorpora questa struttura ma la generalizza alla misura MU, che non presuppone che il destinatario conosca l'intenzione del mittente.

Austin, J. L. (1962). *How to Do Things with Words*. Oxford University Press.

L'introduzione degli atti linguistici. Austin distingue gli atti locutori (il contenuto proposizionale), illocutori (la forza dell'atto: promettere, affermare, ordinare) e perlocutori (l'effetto prodotto nel destinatario). Nel quadro formale, i tre livelli corrispondono a tre livelli della trasduzione: il segnale s (livello locutorio), l'intenzione codificata i_α (livello illocutorio), e l'effetto prodotto nel destinatario $\text{Dec}_\beta(s)$ (livello perlocutorio). Il malinteso può avvenire a ciascuno dei tre livelli indipendentemente dagli altri, e questa stratificazione — che Austin descrive fenomenologicamente — è formalizzata nella monografia attraverso la decomposizione della trasduzione.

Searle, J. R. (1969). *Speech Acts: An Essay in the Philosophy of Language*. Cambridge University Press.

Lo sviluppo sistematico della teoria degli atti linguistici. Searle introduce la distinzione tra intenzione comunicativa e intenzione informativa — il mittente intende che il destinatario riconosca la propria intenzione, non soltanto che riceva il messaggio. Questa struttura intenzionale meta-livello è la condizione perché le implicature funzionino: il destinatario deve riconoscere che il mittente intende comunicare attraverso la deviazione dalla letteralità. La monografia usa questa struttura nel Capitolo I.7 senza adottare la teoria dei background conditions di Searle, che presuppone una comunità di pratiche condivise incompatibile con il quadro del malinteso come condizione strutturale.

II. Analisi della conversazione e riparazione

Schegloff, E. A., Jefferson, G., & Sacks, H. (1977). The preference for self-correction in the organization of repair in conversation. *Language*, 53(2), 361–382.

Il testo fondativo dell'analisi della conversazione sulla riparazione. Schegloff, Jefferson e Sacks documentano sistematicamente le strutture preferenziali della riparazione in conversazioni ordinarie: la preferenza per la riparazione auto-iniziata e auto-eseguita rispetto a quella etero-iniziata ed etero-eseguita, il posizionamento sequenziale delle diverse forme di riparazione, le conseguenze interazionali del mancato avvio della riparazione. Il Capitolo VI formalizza questi risultati nel quadro dell'ottimizzazione

della soglia di riparazione ϵ_{rep} , con la preferenza per l'auto-riparazione come conseguenza del basso costo C_{rep} quando è il mittente stesso a correggere.

Schegloff, E. A. (1992). Repair after next turn: The last structurally provided defense of intersubjectivity in conversation. *American Journal of Sociology*, 97(5), 1295–1345.

Schegloff analizza le sequenze di riparazione posizionate dopo il turno successivo — un tipo di riparazione tardiva che rivela malintesi che erano rimasti latenti per più turni. Questo documento è essenziale per la formalizzazione del malinteso accumulato: la riparazione tardiva dimostra che MU può accumularsi attraverso più turni senza essere rilevata, finché un'incongruenza sufficientemente saliente la porta alla superficie. La soglia ϵ_{rep} non è dunque applicata localmente a ogni turno ma integrata nel tempo — il che spiega perché malintesi piccoli ma sistematici possono cristallizzarsi in narrazioni condivise false.

Levinson, S. C. (1983). *Pragmatics*. Cambridge University Press.

Il riferimento enciclopedico per la pragmatica linguistica. Levinson tratta sistematicamente deissi, presupposizione, implicatura e atti linguistici con un rigore tecnico che la monografia usa come sfondo. Particolarmente rilevante è il capitolo sulla presupposizione: le presupposizioni sono elementi dell'intenzione i_α che il mittente non codifica esplicitamente nel segnale — che dà per scontati nel sistema di rappresentazione del destinatario. Quando M_β non contiene le strutture presupposte, la trasduzione introduce MU_{sys} proporzionale alla densità delle presupposizioni non condivise: il malinteso sistematico cresce con il numero di presupposizioni implicite.

III. Traduzione, fertilità e storia delle idee

Benjamin, W. (1923/1968). The task of the translator. In *Illuminations* (H. Zohn, Trans., pp. 69–82). Schocken Books.

Il testo più ricco filosoficamente sulla traduzione come trasformazione creativa. Benjamin argomenta che la traduzione non è trasmissione del significato ma rivelazione di una "lingua pura" sottostante — una struttura profonda che l'originale e la traduzione condividono e che nessuno dei due esprime completamente. Nel quadro formale, la "lingua pura" di Benjamin corrisponde alla struttura i_α con alta KC che né $Enc_{\{M_\alpha\}}$ né $Enc_{\{M_\beta\}}$ possono codificare completamente: il residuo che la traduzione produce come malinteso creativo. La monografia usa Benjamin come motivazione fenomenologica senza adottare la sua metafisica della lingua.

Steiner, G. (1975). *After Babel: Aspects of Language and Translation*. Oxford University Press.

Il testo più completo sulla traduzione come fenomeno cognitivo e culturale. Steiner propone il modello del processo traduttivo in quattro fasi — fiducia, aggressione, incorporazione, restituzione — che corrispondono nel quadro formale a: la

presupposizione che il segnale contenga i_α (fiducia), la decodifica $Dec_\beta(s)$ che distorce i_α nel sistema M_β (aggressione), l'integrazione di i_β nel sistema M_β (incorporazione), e il tentativo di compensare la distorsione verso i_α (restituzione). La restituzione — l'unica fase che non ha analogia diretta nel quadro formale statico — è il tentativo dinamico di ridurre MU attraverso successive revisioni, analogo alla sequenza di riparazione del Capitolo VI.

Kuhn, T. S. (1962). *The Structure of Scientific Revolutions*. University of Chicago Press.

Già citato nella prima monografia, qui rilevante per l'aspetto specifico dell'incommensurabilità. Kuhn argomenta che i paradigmi scientifici sono incommensurabili — i termini centrali hanno significati diversi nei due paradigmi, e la comunicazione tra paradigmi è sistematicamente distorta. Nel quadro formale, l'incommensurabilità è $MU_{sys}(\text{paradigma A}, \text{paradigma B}) \rightarrow \infty$ per i termini centrali: la trasduzione produce malintesi così profondi da rendere la riparazione convergente impossibile. La transizione tra paradigmi non avviene attraverso la persuasione razionale ma attraverso la sostituzione generazionale — una dinamica che il modello della fertilità delle idee del Capitolo II.4 cattura come catena di trasduzione con alta amplificazione DS.

Pickering, A. (1995). *The Mangle of Practice: Time, Agency, and Science*. University of Chicago Press.

Pickering propone che la pratica scientifica sia un processo di "resistenza e accomodamento" — un dialogo tra le strutture materiali del mondo e le strutture concettuali degli scienziati in cui ciascuna parte modifica l'altra. Nel quadro formale, la resistenza è il segnale di errore predittivo che rivela MU_{intra} tra le aspettative dello scienziato e le osservazioni: è il rilevamento del malinteso intrapersonale tra il sistema di rappresentazione dello scienziato e la struttura del dominio. L'accomodamento è la riparazione — la modifica di M_α nella direzione che riduce MU_{intra} . La monografia usa Pickering come evidenza che il malinteso non è solo un fenomeno intersoggettivo ma caratterizza anche la relazione tra un agente e il dominio che studia.

IV. Potere, ideologia e epistemologia critica

Fricker, M. (2007). *Epistemic Injustice: Power and the Ethics of Knowing*. Oxford University Press.

Il testo che introduce il concetto di ingiustizia epistemica. Fricker distingue l'ingiustizia testimoniale — la deflazione del credito epistemico attribuito a un testimone per ragioni identitarie — dall'ingiustizia ermeneutica — la mancanza di risorse concettuali per comprendere la propria esperienza. Il Capitolo III.7 formalizza l'ingiustizia ermeneutica come costo differenziale $Cost_k(i_\beta)$ elevato per gli agenti il cui M_k è distante da $M_{\{LLM\}}$, e la violenza epistemica come riduzione artificiale di $Acc(i_1)$ che rende il costo strutturalmente insuperabile. La formalizzazione aggiunge al lavoro

di Fricker una struttura quantitativa che permette di distinguere gradi di ingiustizia e di identificarne le cause computazionali.

Bourdieu, P. (1991). *Language and Symbolic Power* (J. B. Thompson, Ed., G. Raymond & M. Adamson, Trans.). Harvard University Press.

Bourdieu analizza il linguaggio come campo di forze simboliche in cui il valore dei discorsi — la loro autorità, la loro credibilità, la loro capacità di produrre effetti — dipende non dal contenuto ma dalla posizione istituzionale del parlante. Nel quadro formale, l'autorità simbolica di Bourdieu corrisponde all'asimmetria dei tassi di aggiornamento $\eta_{\{0i\}}$: i discorsi dei parlanti con alta autorità simbolica hanno $\eta_{\{0i\}}$ alto e orientano la deriva collettiva delle interpretazioni nella propria direzione. La formalizzazione rende esplicita la struttura computazionale del meccanismo che Bourdieu descrive sociologicamente.

Foucault, M. (1969). *L'archéologie du savoir*. Gallimard.

Foucault analizza il discorso come insieme di pratiche che producono e regolano ciò che può essere detto, da chi, in quali contesti — le "formazioni discorsive" che definiscono il campo dei malintesi possibili e impossibili. Nel quadro formale, una formazione discorsiva è un vincolo sulla distribuzione $P_{\{\text{corpus}\}}(M)$: regola quali sistemi di rappresentazione sono ammessi nella produzione del testo che alimenta la prior collettiva, e quindi determina $M_{\{LLM\}}$ per quella comunità. La monografia usa Foucault come sfondo teorico per la politica della memoria del Capitolo V della monografia sulla memoria e per la violenza epistemica del Capitolo III.7.

V. Autoinganno, razionalizzazione e psicologia cognitiva

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, 84(3), 231–259.

Il testo empirico fondamentale sulla razionalizzazione. Nisbett e Wilson documentano sistematicamente che gli esseri umani non hanno accesso diretto ai propri processi cognitivi e producono spiegazioni post hoc che non corrispondono alle cause effettive del loro comportamento. La spiegazione nel quadro formale — $\text{Dec}_{\{M_{\{t_1\}}\}}(\text{Enc}_{\{M_{\{t_0\}}\}}(i)) \neq i$ per sistemi di rappresentazione con gap tra sistema subcoscio e sistema narrativo — non era disponibile a Nisbett e Wilson, che interpretano il fenomeno in termini di inaccessibilità dei processi subcosci senza formalizzarne la struttura. La monografia fornisce la struttura computazionale che Nisbett e Wilson lasciano implicita.

Mele, A. R. (2001). *Self-Deception Unmasked*. Princeton University Press.

L'analisi filosofica più rigorosa dell'autoinganno. Mele argomenta contro la visione "sartriana" dell'autoinganno come paradosso logico — come può l'agente sapere P e credere non-P simultaneamente? — proponendo una teoria deflazionistica in cui

l'autoinganno è il prodotto di bias motivazionali che distorcono la raccolta e l'elaborazione dell'evidenza. Nel quadro formale, i bias motivazionali di Mele corrispondono all'asimmetria di η nel modello del Capitolo VII.4: la riduzione del tasso di aggiornamento verso le interpretazioni che contraddicono la rappresentazione di sé valente produce esattamente la struttura che Mele descrive senza la struttura computazionale che la monografia fornisce.

Freud, S. (1901/1960). *The Psychopathology of Everyday Life* (A. Tyson, Trans.). Norton.

Il testo classico sul lapsus. Freud interpreta il lapsus come produzione involontaria di contenuti repressi — la manifestazione di intenzioni subconscie che "sfuggono" al controllo conscio. Nel quadro formale, il meccanismo è la sovrapposizione $s = (1-\alpha) \cdot s_{\text{int}} + \alpha \cdot s_{\text{sub}} + \epsilon$ del Capitolo VII.5, con i contenuti "repressi" di Freud corrispondenti al segnale subcoscio s_{sub} con peso $\alpha > 0$. La monografia non adotta la teoria della repressione — che presuppone un meccanismo dinamico di mantenimento attivo dei contenuti inconsci — ma usa il quadro formale per captare il fenomeno osservato senza le assunzioni metapsicologiche freudiane.

Bateson, G., Jackson, D. D., Haley, J., & Weakland, J. (1956). Toward a theory of schizophrenia. *Behavioral Science*, 1(4), 251–264.

L'articolo che introduce il concetto di doppio vincolo. Bateson et al. argomenta che l'esposizione prolungata a comunicazioni con struttura di doppio vincolo — messaggi in cui il rispetto di un livello implica la violazione di un altro — produce confusione comunicativa persistente e può contribuire alla patologia schizofrenica. Il Capitolo III.5 formalizza il doppio vincolo come condizione $\nexists i_{\beta}: MU(s_{\text{espl}} \rightarrow i_{\beta}) \leq \epsilon$ e $MU(s_{\text{meta}} \rightarrow i_{\beta}) \leq \epsilon$, mostrando che si tratta di una proprietà strutturale della comunicazione — non di una caratteristica del mittente o del destinatario — che emerge dall'incompatibilità di due vincoli sulla trasduzione.

VI. Bias nei sistemi di linguaggio artificiale

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 610–623).

Il testo che ha introdotto la metafora del "pappagallo stocastico" per i modelli di linguaggio — sistemi che producono testo statisticamente plausibile senza comprensione del significato. Nel quadro formale, il "pappagallo stocastico" è il sistema con $\text{Dec}_{\{\text{LLM}\}}(s) = \text{campione da } p_{\{\text{LLM}\}}(\cdot \mid s)$ senza accesso a M_{α} del mittente — esattamente la struttura del Capitolo IV. La monografia si distanzia da Bender et al. su un punto: la metafora del pappagallo suggerisce che il sistema non faccia nulla di cognitivamente interessante, mentre il quadro formale mostra che esegue una trasduzione strutturata — una media pesata di decodifiche — con proprietà precise e conseguenze prevedibili.

Blodgett, S. L., Barocas, S., Daumé III, H., & Wallach, H. (2020). Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 5454–5476).

La rassegna più completa sul bias nei sistemi di NLP. Blodgett et al. documentano che i termini "bias", "equità" e "danni" nella letteratura NLP sono usati in modo non sistematico, coprendo fenomeni strutturalmente eterogenei. Nel quadro formale, la monografia distingue due fonti di bias che la letteratura spesso confonde: il bias come distanza $MU_{sys}(\alpha, LLM)$ proporzionale alla distanza tra il sistema dell'utente e $M_{\{LLM\}}$ — un bias cognitivo strutturale — e il bias come distorsione semantica $DS(\alpha, LLM)$ nella direzione dei pattern del corpus — un bias contenutistico. Le due misure hanno cause diverse e richiedono rimedi diversi.